Protein surface analysis by Dimension Reduction with applications in functional annotation and drug target prediction

A thesis

Submitted to the Faculty

of

Drexel University

By

Heng Yang

in partial fulfillment for the

requirements of the degree

of

Doctor of philosophy

June 2015

TABLE OF CONTENTS

LIST OF TABLES

LIST OF ILLUSTRATIONS

Abstract

The protein structure initiatives have increased the number of experimentally determined protein tertiary structures, providing tremendous opportunities for detailed comparative analysis of proteins. Although protein structures provide the most exquisite type of molecular information that can yield mechanistic insights into how proteins function, there are still many protein structures with undetermined or poorly defined functions. Functional annotation from protein 3-D structures has attracted many researchers, with most approaches relying on structural superposition against well-characterized proteins. 3-D structure superposition is a complex and computationally demanding problem; forcing most available approaches to only consider backbone atoms for simplicity and efficiency. In this study, we propose protein surface as a more powerful representation of proteins than the traditional backbone or atomic representations. In order to efficiently analyze protein surfaces, we introduce a novel approach to reduce the 3-D surface to a 2-D image map and utilize image registration algorithms to compare these feature-rich images. Whereas the dimension reduction inherently captures the 3-D geometry of the surface patches, we enrich the image map with additional features known to be important for defining molecular activity of the proteins, such as curvature, electrostatic potential, hydrophobicity, and residue conservation. Comparison of these enriched surface maps using image registration methods allows us to find similar surface patches shared between proteins. While the computational challenges remain to scale our approach to study comparisons in the entire set of available protein structures, our novel approach provides unique advantages compared to other structure comparison methods. We show

that our method is able to detect local similarities even when proteins lack a global structure similarity. We also demonstrate the utility of the image maps and their comparisons in functional annotation, and drug target prediction tasks.

Acknowledgement

First and foremost, I want to thank my Ph.D advisor, Professor Ahmet Sacan. It has been an honor to be his Ph.D student. His continuous support, with his vast knowledge helped me in many ways of my research career. Without his encouragement, patience, guidance and help, this dissertation would not be possible.

Besides my advisor, I would like to thank my committee members for their insightful comments and suggestions.

I also want to express my gratitude to my family, who raised me, educated me and give me support throughout my entire life.

I want to thank my lab mates for working together, and especially thank Rehman Qureshi for reviewing and editing my research papers.

I also want to thank my friends, Junyi Xiao, Bochao Zhang who always stood by me through good times and bad. Thank you for listening, offering me advice, and supporting me through this venture.

Last but not the least, I take this opportunity to express my gratitude to Biomedical Engineering department of Drexel University, which have provided me with financial assistance in my Ph.D study.

Chapter 1: Introduction

Protein structures

Protein is composed of a chain of 20 different amino acids. In order to perform the biological function, those amino acids are brought up together to form a specific 3-D conformation by the force of hydrogen bonding, Van der Waals' force, and hydrophobic interaction. Thus, to understand the function of a protein, it is important to look at its 3-D conformation.

The number of solved three-dimensional protein structures available in the Protein Databank (PDB) is increasing rapidly. There are now over 100,000 protein structures in the PDB [1]. This data better enables the study of evolutionary, functional, and structural relationships between proteins. Protein structures capture information not readily available in protein sequences. Two amino acids that are far apart in a sequence could be close together in 3-D due to protein folding. Two protein structures can be aligned in a way analogous to sequence alignment. These alignments provide a measure of the similarity between two protein structures. This is a useful technique because protein structure is better conserved by evolution than protein sequence. Thus, protein structure comparison is able to detect more distant evolutionary relationships than sequence comparison alone. Almost all structural alignment methods deal with the backbone chain of the protein only, which is a gross simplification of the complex shape of the proteins.

In bioinformatics, genomic and structural data is increasingly available. Databases containing gene sequences, protein sequences, and protein structures are growing rapidly.

This deluge of data has allowed the identification of similarities in structures and sequences. This enables the characterization of the similarity of biomolecules resulting from convergent or divergent evolution and the identification of functional similarity.

Determining the functions of individual proteins is essential to understanding their contribution to the behavior of the cell and the organism as a whole and creates tremendous therapeutic opportunities for treating diseases. Availability of large scale genomic and proteomic data has invited development of automated computational methods for functional annotation of proteins. Traditionally, sequence analysis has been the main source of information, where pairwise, multiple alignments and statistical and machine learning methods have been utilized for classification of proteins into known functional families. However, sequence alone becomes insufficient for making functional inferences for distantly related proteins or those proteins that have the same function through convergent evolution, since protein function is primarily related to 3-D shape and structure of the protein. Protein structure is considered as more informative than sequence alone, because proteins function via interactions with ligands and other proteins, placing structure under greater evolutionary pressure than primary sequence information, and making structural similarities between homologous proteins detectable even under low sequence similarity conditions.

Protein data bank (PDB)

Protein data bank is a repository of structural information for large biological molecules. Since PDB has been established in 1971, the number of structures in the repository grows exponentially, and as of today (May/2015), there are 108957 biological macromolecules in the database. Among the number of macromolecular structures in the database, 92.9%

are proteins, 1.4% are DNAs, 1% are RNAs, and 4.6% are mixed structures. For every five years, there are on average 30000 structures deposited into the database, the number for each year deposit is also growing rapidly. PDB provides SCOP classification for details of structural and evolutionary hints. Understanding the structure and shape of a molecule is very important in understanding a certain disease and drug development. However, many proteins deposited into PDB by structural genomics projects are lack function annotation, and the problem comes to how to deduce protein functions from the structures in absence of function information. Much work have been done for molecular recognition by using computational methods, and various approaches have been proposed for ligand prediction and function annotation. However, current computational tools still lack of accuracy in recognizing similar binding sites among proteins and it makes it clear that this problem is way more complicated and difficult than was expected. The current computation methods for protein structure analysis are introduced in the next section.

For the latest PDB information, please refer to http://www.rcsb.org/pdb/home/home.do.

Protein structure alignment

Corroborating the importance of structural data, protein structure initiatives have been established with the goal of expanding the repository of experimentally determined protein structures. As many as 26% of the structures resulting from these structural genomics initiatives have unknown or putative function [2]. Consequently, numerous approaches have been developed for comparing and data mining protein structures, with the hopes of finding functionally relevant similarities among both well-studied and less characterized proteins.

Structure alignment methods make up a majority of the available structure comparison approaches. In structure alignment, one seeks to find correspondences between the residues of the proteins being compared and also a translation/rotation matrix that best superposes these corresponding residues. Finding an optimum structural alignment is computationally difficult and available methods employ heuristics to find near-optimal solutions within practical execution times. Available methods are based on distance matrices[3], common subgraph searches[4], geometric hashing techniques[5], genetic algorithms[6], and Delaunay tessellations[7]. An important drawback of structure alignment methods has been their prudent use of geometric information, and only recently methods have been proposed to additionally utilize biochemical and evolutionary information.

Structure alignment methods generally represent each amino acid as a single point space, often using the coordinates of its alpha carbon atom. While this simplification is sufficient for fold recognition purposes where the focus is on categorization of the overall shape of protein domains, it may fail to detect important local arrangements of amino acid side chain atoms. Furthermore, global structural similarity does not necessitate the same enzymatic activity or binding interactions. TIM barrel family of proteins provide an extreme example of this, where proteins sharing the same structural fold can have diverse functions [8]. Ser-His-Asp catalytic triad of serine protease family provide an example from the other extreme, where due to chemical constraints of enzymatic activity, proteins share highly conserved arrangement of active site residues, while having dissimilar global structures [9].

The need to identify conserved local arrangements of a few amino acids, regardless of the overall fold, has motivated development of a new class of methods for discovery and search of structural patterns, such as SPRITE and ASSAM[10], PDBemotif [11], LFM-pro [12], and PROMOTIF [13]. These methods try to find spatial configuration of amino acid residues with well-conserved inter-residue distances and in line with their focus on function rather than structure, they often utilize functional side chain atoms instead of backbone atoms.

It has been observed that proteins with similar active sites have similar functions[14] and that active sites are usually located within pockets formed on the protein surface[15], [16]. These observations have prompted focus on analysis of surface pockets for identification of ligand binding and protein function. Surface pockets have been defined as regions of favorable interact energies [17] or from purely geometric characteristics [18]. Consequently, a class of structure comparison methods have been developed to compare these surface regions.

The methods that make use of only geometric information for comparison of proteins surfaces include those that summarize the shape by descriptors such as Zernike moments [19] and distance-based features [20] and those that represent surfaces as point clouds [21]. The methods based on shape descriptors generally solve the global structure similarity problem, similar to structure alignment methods, whereas the point cloud methods try to detect local residue or atomic arrangements. While existing approaches have been useful in comparing known protein active sites, they are limited in their ability to discover previously unknown functional sites.

The studies by both Fanning et.al. [22] and Pawlowski and Godzik [23] have borrowed ideas from cartography studies and have applied similar projections used therein. Fanning [22] generated a contour map of the surface, in order to preserve some of the topographic features of the irregular protein shapes. They have used Mercator-like projection and Mollweide projection in order to investigate whether topographic features can provide antigenic determinants. Pawlowski and Godzik have used an equal area sinusoidal cartographic projection (also known as the Mercator equal-area projection) as a simple surface representation to measure the map similarity of proteins. These early studies in molecular cartography have remained isolated and the follow up researches are not sufficient. With the current study, we hope to bring the power and appeal of molecular cartography back to the attention of the protein structural analysis community.

In this study, we describe a novel representation and comparison method for protein surface analysis that is able to capture various surface features in a computationally efficient manner. Specifically, we unfold protein surfaces into two dimensional images and perform comparisons using these images. The two dimensional image representation allows the use of faster image registration methods, as opposed to the more demanding graph matching methods required in other methods.

Specific Aims

Proteins generally interact with other proteins and molecules via their surface regions, and a backbone-only analysis of protein structures may miss many of the functional and evolutionary features. Surface information can help better elucidate protein functions and their interactions with other proteins. Computational analysis and comparison of protein

surfaces is an important challenge to overcome to enable efficient and accurate functional characterization of proteins.

In this study we present a new method for protein function prediction using protein surface analysis. The key innovation in this project is to unfold and decompose 3-D Tertiary structure into 2-D maps and encode geometric and biochemical features of the protein, such as hydrophobicity, electrostatic potential, evolution information and curvature, into the 2-D map to capture functionally relevant information. Enriched images can then be compared using efficient 2-D image registration methods to identify surface regions and features shared by proteins. Furthermore, we established the "surface pharmacophore" for drug targeting prediction. Pharmacophores hold exclusive features to distinguish from one another. In this way, screening method can be applied on a pre-built database and hits that are compatible with the target protein are obtained.

Aim 1) Development of a protein surface comparison method

Surface comparison attempts to characterize and compare geometrical and physicochemical features on protein surfaces. We hypothesize that a surface analysis will yield a better predictor of protein function and interactions than sequence or structural information. This first specific aim is to develop a surface comparison method based on mapping 3-D protein surfaces onto the 2-D space, through dimension reduction methods and enriching the 2-D representation with physicochemical and geometrical features. The 2-D representation allows analysis of the surfaces in a computationally feasible manner, and opens the possibility of using numerous image processing methods for surface segmentation and image registration. The application is able to identify similarity of two binding site sections.

Aim 2) Identification of similar surface patches

Protein binding sites serve as signatures for an enzymatic function, protein-protein interaction, and drug binding. The second specific aim pursues the first aim further. It develops the application that, given two proteins, one with known binding site information, is able to locate the binding site of another protein from the information of the known protein.

Aim 3) Prediction of off-Target drug-protein interactions.

An accurate and efficient method for protein surface characterization and comparison can play an important role in rational drug design. For example, analysis of protein surfaces could help identify protein binding pockets so that the requirements for a given pharmaceutical compound's size and binding orientation can be determined. Furthermore, knowledge of the protein conformation helps researchers develop specific pharmaceuticals for a given disease. This analysis can also assist in the investigation of protein-protein interactions and give researchers insight into the biological processes of the cell. We propose to develop an off-target identification system that relies on the 2-D surface signature model. For a given drug with unknown binding site, we will search for 2-D surface signatures, and rank the results by their similarity scores. An advantage of our proposed method over other existing off-target prediction methods is its ability to not only predict potential targets, but also identify the binding sites.

Although this project focuses on drug-binding applications, the representation and the database of protein surface signatures will become useful in other related structural bioinformatics applications, including functional and evolutionary annotation of proteins, prediction of protein-protein interactions, and homology modeling.

Chapter 2: Protein Surface Preliminaries

Protein ligand binding

The protein ligand binding, or the drug receptor interactions is an important study in biochemistry and pharmacology since it leads to the discovery of new drug targets for disease treatment. The measurement of the binding is defined as affinity and efficacy. The affinity is the relationship between a drug and a receptor. If a drug associates closely with a receptor, it has a high affinity to the receptor, and vice versa. Efficacy describes the capability of a drug to alter a receptor and induce the physiological response.

Receptors are macromolecules, such as proteins and DNAs, which regulate cellular biochemical processes between or within cells. Ligands are small molecules that bind to the receptor, alter the conformation of the receptor and eventually activate or inactive the receptor. The interaction between ligands and receptors is through binding, and the binding is determined by a combination of shape complementarity and energetically favorable interactions. Shape complementarity is an essential feature for ligand binding, since proteins have a unique binding site that will only bind a certain ligands (shown in Figure 1), and thus, competitive ligands usually bear a resemblance on the binding region. Protein conformation is dynamic, and chemical forces between protein and ligand, such as electrostatic potential, hydrogen binding, and Van der Waals forces affect the interaction.

There are two ligand binding theories, one is conformation selection and the other is induced fit. Protein ligand interactions involve with conformational change, and conformational selection claims this change happens before the association of the ligand.

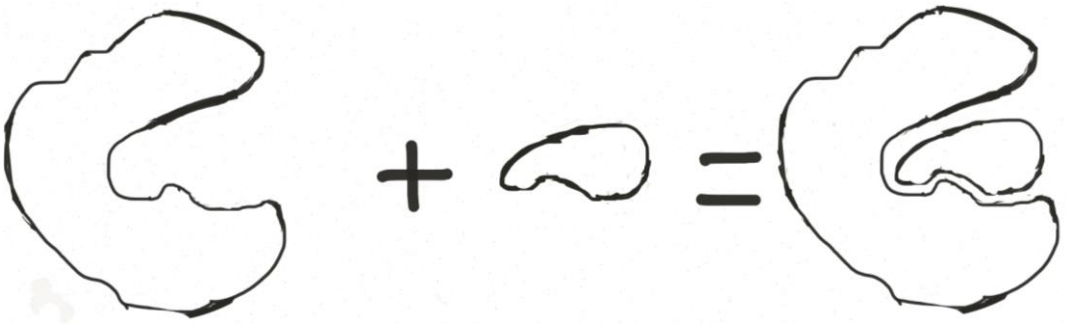

Figure 1 The surface shape complementarity to the binding ligand.

Protein usually adapts a certain conformation in a system that requires the minimum free energy. When a ligand enters into the system, the energy equilibrium is interrupted, and protein alters conformation in response to the new energy of the system, and this alternation favors the binding of the ligand. Inactive receptors become active, and the conformation changes to a new equilibrium.

In contrast, induced fit claims the conformational change happens after the ligand binding. Binding site adjusts the shape to better fit the binding ligand. The surface of the protein at the binding region exhibits a complementary shape to the binding ligand in order to best fit to the ligand shape.

Principles of Protein Surfaces

A number of different representations have been developed to describe the protein surface. A classic representation is the solvent accessible surface, introduced by Lee and Richards [24]. The accessible surface can be defined by simulating a probe "rolling" on the surface. The path traced out by the center of the probe forms the solvent accessible

surface (See Figure 2). Several variations of the protein surface have been defined. Connolly introduced another representation of the protein surface, called the Connolly surface (also known as the solvent-excluded surface) [25, 26] Compared to the solvent accessible surface which is considered the expanded van der Waals surface of the protein, the Connolly surface is defined as the inward-facing part of the probe with the other atoms and their neighbors.

Protein surfaces generated by these methods have found use in a variety of visualization and analysis applications. Almost all popular macromolecular visualization programs now contain routines for the generation and visualization of different types of surface representations [27], [28]. The quantification of the surface area has enabled several discoveries in areas such as protein folding and protein docking sites. Lee and Richards [24] have found the decrease in accessible areas of hydrophobic atoms to be greater than that of hydrophilic atoms, supporting the hydrophobic burial hypothesis in protein folding. The complementarity of surfaces has been exploited in the molecular docking applications that form an essential component of modern rational drug design work flow [29]. Comparative analyses of surfaces have been utilized in the functional annotation of proteins [30].

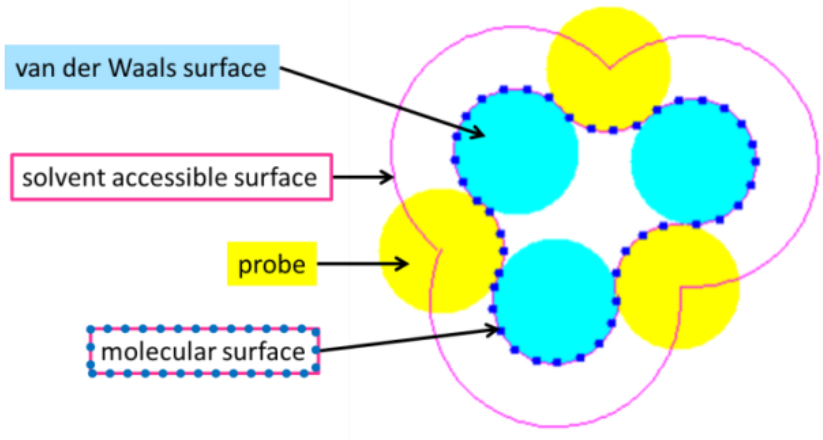

Figure 2 Illustration of solvent accessible and molecular surfaces. The cyan balls represent the van der Waals surface of the atoms. The yellow probe is the ball that is rolled around on the molecule. The path taken by the center of the probe is the solvent accessible surface, shown in pink outline. The surface traced by the probe facing the inside of the molecule is the molecular surface, shown in blue-dotted outline.

3-D Surface and its applications

Connolly developed a numerical algorithm to calculate the 3-D protein contour based on solvent-accessible surface method [31]. Later a surface triangulation method was developed by Connolly [32] which is based on subdividing the curved faces of an analytical molecular surface representation.

Connolly later introduced another representation of a protein surface, called the Molecular surface (also known as the solvent-excluded surface or Connolly surface) [25]. Unlike the solvent accessible surface, which is considered the expanded van der Waals surface of the protein, the Molecular surface is defined as the inward-facing part of the probe that is rolling on the protein surface. In the present study, we utilize this surface representation of the proteins.

Starting from Connolly's work, numerous methods have been proposed for surface representation. Sanner [33] introduced the idea of r-reduced surface and developed an efficient algorithm to compute the outer components of the surface. Staib [34] developed a mathematical surface representation by expansions of spherical harmonic functions, which can be used in analyzing surface curvatures, surface interaction, and surface visualization. While we have utilized the solvent excluded representation in this study, the approach introduced here can be extended to these other surface representations.

Protein surfaces generated by these methods have found use in a variety of visualization and analysis applications. Almost all popular macromolecular visualization programs now contain routines for the generation and visualization of different types of surface representations [27, 28].

The difficulty of dealing with surfaces is apparent, in comparison to the more widely utilized primary sequence or backbone conformation, which possess numerous alignment methods. Due to the complexity of the surfaces and the lack of established methods for general-purpose analysis, most studies have focused on certain surface features, such as active or functional sites and structural motifs [35]. These sites are identified only around a local spatial proximity or surface patch and involve only a few highly conserved amino acids [36].

Approaches that have attempted to represent and analyze the entire surface have been geared toward extracting generic shape parameters that are not amenable to detailed characterization of surfaces. 3-D Spherical harmonics and Zernike descriptors have been used as feature vectors for protein structure comparison and similarity-based retrieval [37]. Geometric hashing has also been used for translation and rotation invariant

comparison of sets of atoms [38]. Poirette [36] has used the genetic algorithm to compare two protein surfaces by searching for a translation and rotation matrix that brings the two surfaces together, maximizing the surface overlap. However, geometrical only methods lack of physicochemical information that are essential to identify the uniqueness of a molecule, like electrostatic potential and hydrophobicity.

Protein Surface features

There are many ways to classify amino acids types, and one common way is by looking at the properties of the side chains. Each amino acid has its specific characteristics defined by the side chain, which gives it chemical properties, such as hydrophobic or hydrophilic.

Feature 1 electrostatic potential

The Coulomb's law states that the electrical force along the straight line in between two charges at rest is directly proportional to the product of the charges and inversely proportional to the square of the distance between them.

Electrostatic force determines on a large part in protein fold, conformational stability, protein - ligand binding as well as protein - protein interactions. The reason behind it is a mixture of positive protons and negative electrons attracting and repelling with this great force. The electrical force that holds the atoms and molecules together, mainly because balance of charge of the interaction region is not perfect, or the distances are very small[39].

Feature 2 hydrophobicity

Hydrophobicity describes the ability of repellent for a molecule to the water, which attributes to the residue properties of a molecule. Most proteins have hydrophobic amino acid residues buried inside, and polar (hydrophilic) amino acids cover the molecule

surface, and interact with water to form hydrogen bonds, which keeps the stability of the molecule. Hydrophobicity plays a central role in determining the overall conformation of a protein. A number of different hydrophobicity scales have been proposed, and in this work, Kyte and Doolittle's method is adopted [40]. The basic algorithm behind Kyte and Doolittle's method is that, each amino acid is given a hydrophobicity score between 4.5 and -4.5. The higher, the more hydrophobic, and the lower, the more hydrophilic. A window contains 9 amino acid sides from the beginning to the end of an amino acid chain, and each time, calculates the average of hydrophobicity score within the window and assigns the average score to the first amino acid in the window.

Feature 3 conservation

In the evolution of proteins, some amino acids tolerate mutations across homologues. The highly conserved amino acids are crucial for function of a protein and often indicate structural relevance and functional importance. Thus, evolutionary conservation analysis of a protein helps uncover regions that are conserved among homologous proteins. The way the evolution conservation is calculated involves three step, first step is to perform a PST-BLAST search against the NCBI non-redundant protein sequence database, and second step is to multiple align the search results using MUSCLE [41], and last step is to derive sequence conservation score for each residue using the method described in STACCATO [42].

Feature 4 curvature

Curvature is a mathematical concept that describes the geometric feature of a structure, and in three dimension, it measures how much the surface bends at a surface point. The tangent drawn along the curve forms the angle at a given point with the curve, and the rate of change of this angle is considered as curvature at that point. If a normal vector is

put perpendicularly to the tangent plane at a point (x0,y0) on the surface z = f(x,y), the normal vector at this point are defined. Normal planes are planes containing the normal vector, and normal plane intersect the surface to form a curve, and the curvature of this curve is called normal curvature. The maximum k1 and minimum k2 of the normal curvatures at a point are principal curvatures. The quantity K = k1*k2 is Gaussian curvature and the quantity H = (k1 + k2)/2 is the mean curvature, those two curvatures play a very important role in the theory of surfaces.

Other features

Features in our application are not limited. There are other important features can be incorporated into the surface enrichment. Residue interface propensity, is the property that measures how likely a residue locates on the interface. Amino acid planarity can also be calculated and enriched in protein surface. Pocket detection algorithms can also be incorporated in our application.

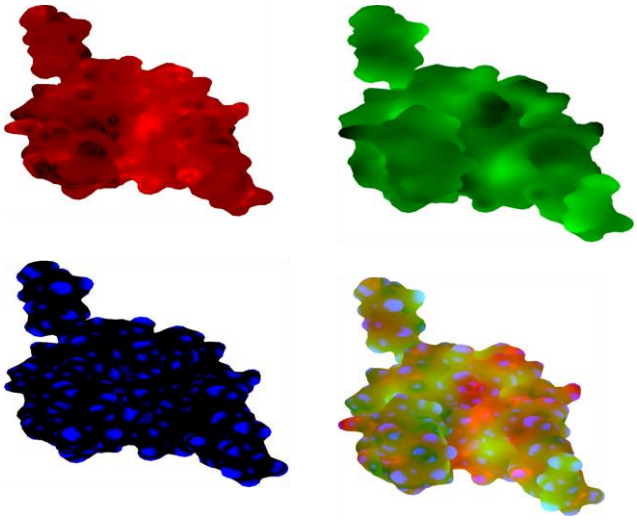

Figure 3 Visualizations of protein (pdb: 1ayl) and protein (pdb: 2yiw). Left upper image is the electrostatic mapping, right upper image is hydrophobicity mapping, left bottom is curvature mapping, and the right bottom is the mapping of three properties.

Figure 3 shows a protein surface with color enrichment. Electrostatic potential is filled in the red channel at the red image, and green and blue channels are left blank. Hydrophobicity is filled in the green channel at the green image and red and blue channels are left blank. Curvature value is filled in the blue channel at the blue image and red and green channels are left blank. The last colorful image is drawn by combining the three features into RGB channels respectively. Regions high or low in these properties are still visually discernible when the three channels are combined into a single colored surface.

Protein Surface alignment

Similar to the conjecture that protein structure provides better information than sequence, it has been suggested that the surface of the protein may provide information not detectable from the structure [43]. Two proteins may have different backbones and

different overall 3-D structures, but may contain highly similar surface regions, giving them the ability to catalyze chemically equivalent reactions on similar substrates [44]. Proteins that meet these conditions are likely experiencing convergent or divergent evolution. In the case of divergent evolution, two protein sequences or structures can mutate over time, but the surface characteristics must be conserved in order to maintain the specific function. In the case of convergent evolution, proteins with similar functions but different structures can evolve similar surface characteristics, causing non-homologous proteins to share similar active or binding sites [45, 46]. The conservation of similar sites on protein surfaces may not be detected by sequence or structure comparison, but the surface determinants can determine the common functionality, making surface based methods invaluable for protein functional annotation. Proteins generally interact with other molecules via their surface regions. The backbone-only analysis may miss many of the functional and evolutionary features. Surface information can help better elucidate protein functions and their interactions with other proteins. Computational analysis and comparison of protein surfaces is an important challenge to overcome to enable efficient and accurate functional characterization of proteins.

Two proteins may have different backbones and different overall 3-D structures, but may contain highly similar surface regions, giving them the ability to catalyze chemically equivalent reactions on similar substrates. The conservation of similar sites on protein surfaces may not be detected by sequence or structure comparison, but the surface determinants can determine the common functionality, making surface based methods invaluable for protein functional annotation. Proteins are assumed to perform similar

functions if they share similar binding patterns, and by comparing surface similarity, one is able to infer protein function.

There are three situations for protein surface comparison. First is comparison between two binding sites, and this is the situation where binding sites of two proteins are known, but one protein with unknown functions. By comparing two binding sites, one can infer the function for the unknown protein. Second is, given a protein surface with unknown binding site, search it in the database to find the most similar binding site. It is in the situation where a binding site database is constructed, and a query protein is searched against the database to look for the binding site similarity. The selected binding site shares the most similar pattern with the query protein, and the binding site on the query protein can be inferred. Third is comparison between two protein surfaces with unknown binding sites, and this is in the situation where the similar patches on proteins are of interest.

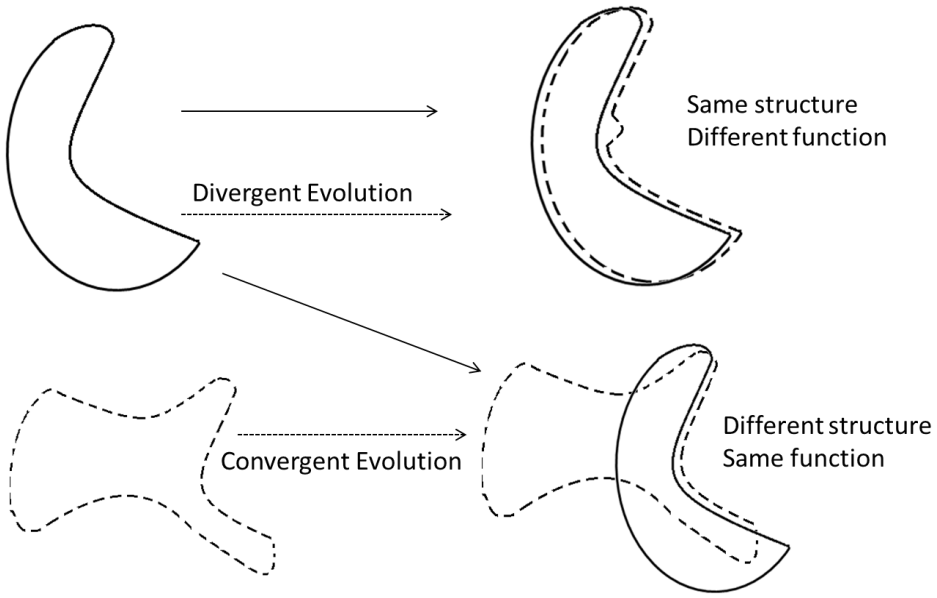

Figure 4 Schematic illustration of importance of local surface characterization compared to structure comparison. In the left picture, two proteins that have almost the same overall structure are shown. Despite the highly similar structures, a small local difference in their binding site due to divergent evolution may cause these proteins to have different functions. In the lower right-hand picture, two structurally different proteins are shown in bold line and dotted line, respectively. Although these proteins differ in their global structure, by convergent evolution, they may share similar local binding sites and may have similar functions.

Advantages of protein surface alignment

Protein surface may provide information not detectable from the structure alone. Two proteins may have different backbones and different overall 3-D structures, but may contain highly similar surface regions, giving them the ability to catalyze chemically equivalent reactions on similar substrates. Proteins that meet these conditions are likely experiencing convergent or divergent evolution. In the case of divergent evolution, two protein sequences or structures can mutate over time, but the surface characteristics must be conserved in order to maintain the specific function. In the case of convergent

evolution, proteins with similar functions but different structures can evolve similar surface characteristics, causing non-homologous proteins to share similar active or binding sites, see Figure 4.

An accurate and efficient method for protein surface characterization and comparison can play an important role in rational drug design. For example, analysis of protein surfaces could help identify protein binding pockets so that the requirements for a given pharmaceutical compound's size and binding orientation can be determined. Furthermore, knowledge of the protein conformation helps researchers to develop specific pharmaceuticals for a given disease. This analysis can also assist in the investigation of protein-protein interactions and give researchers insight into the biological processes of the cell. For example, signal transduction is carried out by a cascade of protein-protein interactions. Moreover, the ligand binding sites act as a signal trigger located on the protein surface. Once, the ligand binds to the protein's active site, it alters the protein's 3-D structure and thus triggers a certain response.

Chapter 3: Dimension Reduction

Dimension Reduction Introduction

A Dimension Reduction Method (DRM) is a geometric technique that collapses higher dimensional data into lower dimension space by extracting maximum variance among data points. There are many different DRMs, but they generally fall into three categories, linear methods, global nonlinear methods, and local nonlinear methods. Principal Components Analysis (PCA) is one of the classic linear dimension reduction methods. It attempts to find a linear mapping between high dimensional and low dimensional data using the principal eigenvectors of the covariance matrix of data. PCA calculates a projection direction of the data which represents the best of the original data. However, since the principal eigenvectors rely mainly on the data dimensionality, PCA is not good at reducing relatively high dimensional data. Linear discriminant analysis (LDA) calculates projections that maximize the discrimination of disparate classes in the data, which enables its usage as a linear classifier. LDA attempts to project higher dimensional data onto a hyperplane that maximizes the separability between two groups. Locally Linear Coordination (LLC) and Coordinated Factor Analysis (CFA) are the other linear DRMs we investigated [47].

Multidimensional Scaling (MDS) is a global nonlinear DRM. MDS constructs a dissimilarity matrix in high dimensional data points using, for example, geodesic distance in three dimension, and utilizes a stress function to measure the pairwise distance between the corresponding data points in high dimensional (geodesic distance) and low dimensional (Euclidean distance) and tries to maintain the minimum distance errors.

Commonly used stress functions are Kruskal's stress and Sammon's stress functions [47]. When Sammon's stress function is used, the method is often called Sammon Projection or Sammon Mapping. Other global nonlinear DRMs that we investigated in this study include Stochastic Neighbor Embedding (SNE), Isomap, and Stochastic Proximity Embedding (SPE). Like MDS, SNE attempts to preserve pairwise distances between data points in low dimensions, but it defines a new distance measure and a corresponding error function, such that local properties of a manifold are better preserved. tSNE is T-distributed Stochastic Neighbor Embedding, which is an extension of SNE that provides a more efficient calculation of the error function. The Isomap method addresses the surface manifold problem, where two points with a small Euclidean distance may be far apart on the manifold, by utilizing shortest paths on a graph of nearest neighbors. Isomap is able to unroll the manifold and maintain the point relationship on 2-D. Figure 5 and Figure 6 shows the results of the unfolded surface manifold. Each color strip appears on 3-D is very well preserved on 2-D, the point relationship is kept after dimension reduction from Isomap.

Another global non-linear method is the Multilayer Autoencoder, which is a multi-layer feed-forward neural network. When the network is trained on data, the middle hidden layer contains a lower dimensional representation of the data points that preserves as much of the original data as possible [47].

We have investigated several local nonlinear DRMs including Locally-Linear Embedding (LLE) and local Tangent Space Analysis (LTSA). LLE involves finding the nearest neighbors of each point and then determining weights for each point in order to express the point as a linear combination of its neighbors. The weights are a set describing how

much each neighbor contributes to determining the location of the given point. LLE then uses the set of weights to place the point in a lower dimensional space. LLE is faster than Isomap when it uses sparse matrix algorithms, but cannot handle non-uniform sample densities very well. The LTSA method uses the local tangent space of each data point to describe local properties of high dimensional data. LTSA assumes that there is a linear mapping from a data point in higher dimensions and the corresponding point in low-dimensional space to the same local tangent space. LTSA simultaneously searches for the coordinates of data points in lower dimensions and for the mappings to the local tangent space of the high-dimensional data[47].

Our application adopts Isomap as our dimension reduction method in that protein is a complex structure, and Isomap calculates the geodesic relationship among points, which is able to capture the actual neighborhood information of points and unfold the structure.

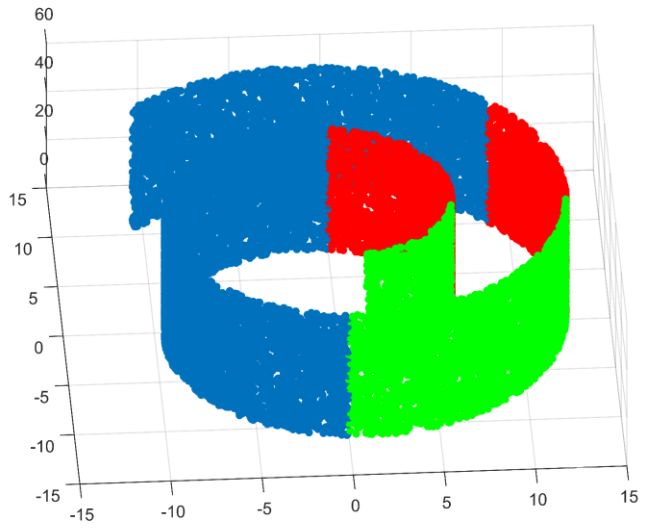

Figure 5 3-D surface manifold.

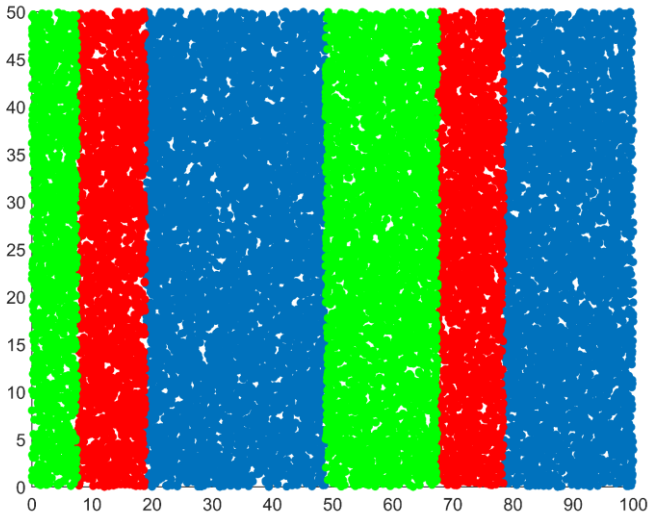

Figure 6 2-D points of surface manifold using Isomap.

Structure segmentation

Our initial solution was to map the enclosed 3-D structure onto the 2-D image, however, this solution has a major drawback, which causes the overlapping of triangle edges. This is due to the inherent property of an enclosed 3-D surface. It is not possible to equally maintain geometric relationship of all the surface points. Notably, the points in 2-D would have different local neighbors than they had in 3-D. In order to alleviate this problem, we section an enclosed surface and consider each sub-surface separately. There are two different ways for the structure segmentation: General segmentation and binding site segmentation. General segmentation cuts a structure into multiple different sub structures. The segmentation algorithm ensures the radius of each section is at most 15

angstrom, and it also defines an exclusion radius that all the points within this radius in one section will not be selected again for another section to avoid generating identical section. The default value of exclusion radius is 10 angstrom. However, points that locate beyond exclusion radius can be re-selected by other sections to ensure some section overlap. These two radius parameters guarantee that the binding site region of a protein will be contained fully in at least one section. This segmentation results in different number of protein sections based on the size of a protein. The binding site segmentation assumes known binding site points beforehand, and it only cuts the binding site section according to the location of binding site points. The binding site points are calculated by comparing Euclidean distance between surface points and ligand. The ligand coordinates are fetched from PDB file. The closest surface points to the ligand are considered as binding site points. The binding site center and the center of the 3-D structure are computed and connected with surface points. The angles formed between surface points and two centers are used to segment the binding site section: points with angles less than 180 degree are considered as the part of binding site section, whereas other points are excluded. Figure 7 shows binding site segmentation (upper row) and general segmentation (bottom row) with color enrichment. For the upper row where the binding site location is known, only one sub structure is segmented, which contains the full binding site region. The third image is a 2-D image generated from the sub structure using Isomap. Since the binding site location is known, the binding site region on 2-D images easily deduced. When the binding site location is unknown, the general segmentation is applied. Bottom row shows one of multiple segmentations, and each segmentation results in a 2-D image.

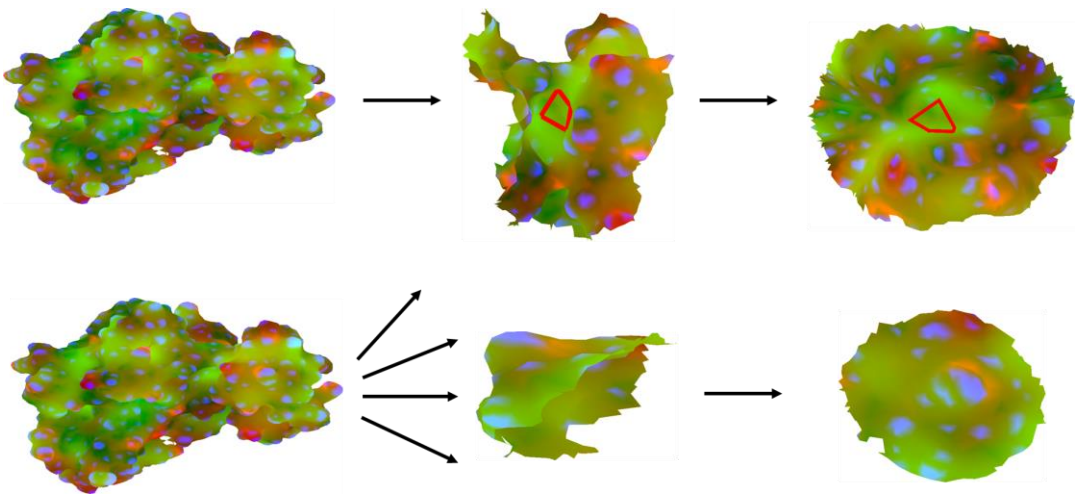

Figure 7 Binding site segmentation and general segmentation. Binding site region is shown in red polygon.

Dimension reduction evaluation

The success of the proposed approach can be measured by its ability to detect similarities between known active sites in related proteins. While we leave a large scale evaluation of the proposed approach as a future work, we describe here the evaluation and optimization of the dimension reduction component of the proposed approach.

For a given dataset of protein structures, we evaluate each dimension reduction method based on its accuracy and speed in mapping the surface. The accuracy of the dimension reduction method is evaluated as the ability to preserve the spatial features and relationships among the points. For the purpose of dimension reduction evaluation, the general segmentation only cuts the structure into 6 sections according to the XYZ plane,

in other words, we select surface points according to their XYZ coordinates. Figure 8 shows a comparison of segmentation between an enclosed sphere and a semi-sphere that cut with our method, PCA is used to obtain a 2-D map. The comparison clearly shows that PCA is able to better preserve the local geometric properties of the surface points on an open structure rather than an enclosed structure.

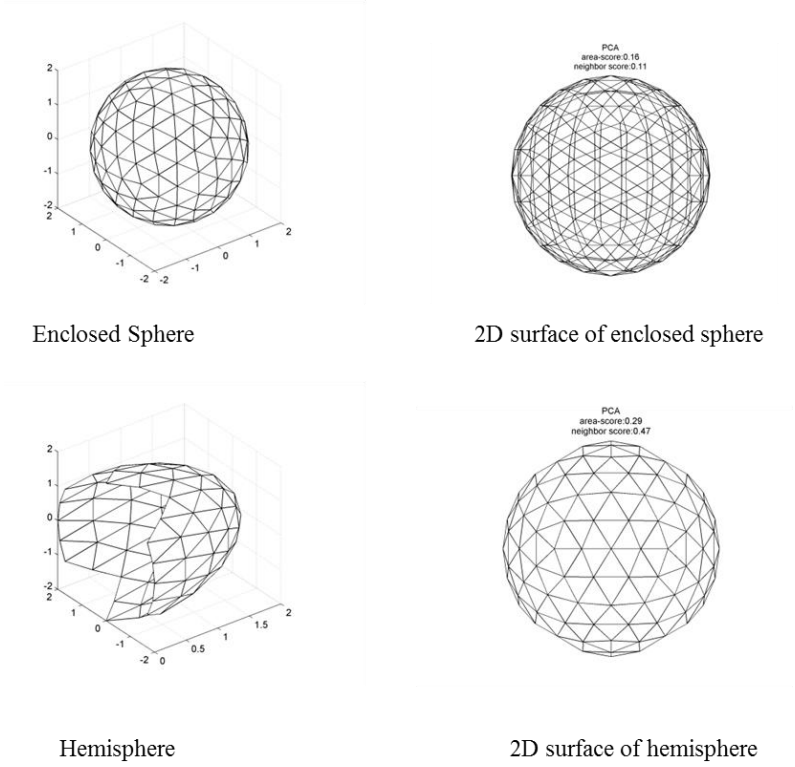

Figure 8 Different results of DRMs between Enclosed Sphere and Hemisphere.

Two assessment criteria are defined: Area Score ranging from -1 to 1 and Neighbor score ranging from 0 to 1. The Area Score is calculated as the Pearson's correlation coefficient for the areas of the triangles in 3-D and in 2-D. The Area score acts as a measure of

unequal distortions induced by the mapping procedure. The higher the correlation value, the better is a method in preserving relative spatial distributions of the points. The Neighbor Score evaluates the ability of a method to preserve the neighborhood relationships among points, and is calculated using Tanimoto similarity coefficient of the connectivity matrices of all points in 2-D and 3-D. In 3-D, the connectivity matrix is obtained using k-nearest neighbors of each point using their geodesic distances. In 2-D, the connectivity matrix is obtained using k-nearest neighbors of each point using the Euclidean distance. These connectivity matrices are then represented as linear bit vectors (with only 0 or 1 values) X and Y. The higher the Tanimoto coefficient the better the method is at preserving neighbors of points.

Dimension reduction methods can be classified into four groups: Traditional linear group, Local non-nonlinear group, Global nonlinear and Global linear group. The evaluation for each group is represented in Figure 9, Figure 10 and Figure 11. An evaluation for each method is conducted using a hemisphere with 2562 points and 5120 triangles. Table 1 shows time and accuracy of each of the Dimension Reduction methods.

Table 1 Performance of Dimension Reduction Methods

Area	KNN	Methods	Neighbor	Runtime(min)
0.28859	3	PCA	0.46901	0.0177
0.28636	3	LLE	0.47205	0.058

0.1189	3	Laplacian	0.29508	0.025
-0.0039	3	LLC	0.021552	0.093
0.017331	3	AutoencoderEA	0.11268	0.344
0.72547	3	SNE	0.53896	13.08
0.030623	3	SymSNE	0.0186	10.42
0.27105	3	CFA	0.123	1.95
0.28859	3	GPLVM	0.469	1.48
0.2885	3	NPE	0.4751	0.045
-0.00774	3	LPP	0.1791	0.02
0.28859	3	LLTSA	0.46901	0.05
0.28859	3	NCA	0.095532	3.1
0.28859	3	MCML	0.46901	77.9
0.28859	3	LDA	0.37524	0.02
0.24604	3	FactorAnalysis	0.20713	0.02
0.16451	3	tSNE	0.53564	3.18
0.81525	3	Isomap	0.53896	5.87
0.69545	3	LandmarkIsomap	0.53564	1.32

0.28859	3	ProbPCA	0.48125	0.06
0.04168	3	KernelPCA	0.23868	0.70
0.28859	3	MDS	0.46901	0.0
0.30813	3	DiffusionMaps	0.48125	0.15
0.73736	3	Sammon	0.54565	4.74
0.10401	3	Sinusoidalcartography	0.45102	0.017

The first chart illustrates the area score, some of the Global nonlinear methods perform better on preserving the area similarity between 3-D space and 2-D space, however, there are still some global methods do not show a good ability on the area consistency.

The second bar chart is the neighbor score analysis. Both the local nonlinear and global nonlinear methods show a strong capability of keeping the same neighbors. The SNE's neighbor score is approximately equal to Sammon, but its running time is almost 3 times slower than Sammon. Isomap and Landmark have an overall better performance than tSNE due to their both good neighbor preservation and faster speed.

The last bar chart is the running time performance. The MCML method is slowest among the methods, and there are two global nonlinear methods (SNE takes 13 minutes and SymSNE takes 10.42 minutes) take a relatively longer time to finish. Based on the performance, LLC, NCA, AutoEncoderEA, symSNE, CFA, might be excluded from the

further investigation due to their unsatisfactory performance. The chart also indicates that area score positively correlates with neighbor score, indicating the failure of maintaining the same area is due to the triangle distortion where the neighbor points have been moved away.

Our approach attempts to address more complex shapes, as are present in almost all proteins. Isomap [48] exploits geodesic distance instead of straight-line Euclidean distance for nonlinear dimension reduction. By using geodesic distances, Isomap preserves intrinsic geometry of the manifold and avoids the "Swiss roll" problem where points far apart in the manifold are deceptively mapped close to each other on lower dimension. In order to obtain the overall color for each triangles in 2-D, we interpolate the vertex color across each triangle on the surface.

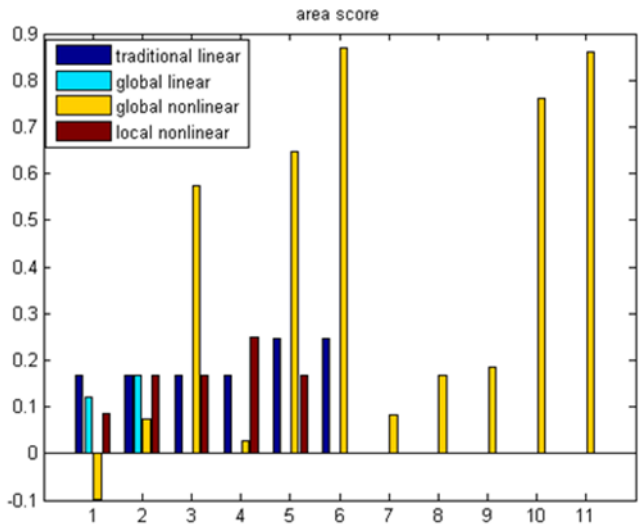

Figure 9 Area score for different dimension reduction methods.

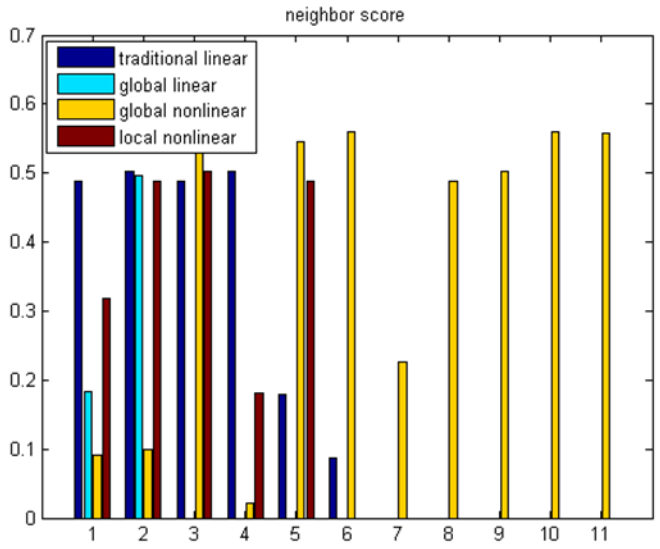

Figure 10 Neighbor score for different dimension reduction methods.

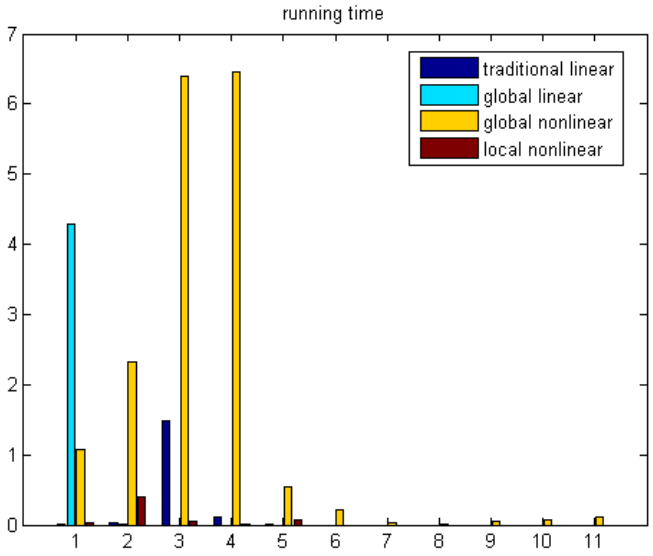

Figure 11 Running time for different dimension reduction methods.

Dimension reduction sample results

Dimension reduction methods handle different problems. For a specific problem, different methods may have complete different results. Figure 12 shows dimension reduction results for a 3-D point cloud of a human face from 6 different methods, the human face mesh is computed using [49]. The first row shows 3 methods that perform relatively well since they are able to preserve the neighbor relations and the human face is still recognizable. Isomap performs the best among the three since fewer triangle overlap occurs in the 2-D mesh compared to PCA and LLE. It is also clear to see Isomap tries to unwrap the 3-D structure instead of simply projecting it on a plane. The second row shows three methods that are not able to handle face structure. They are neither able to keep the neighbor relations nor keeping the original triangle areas.

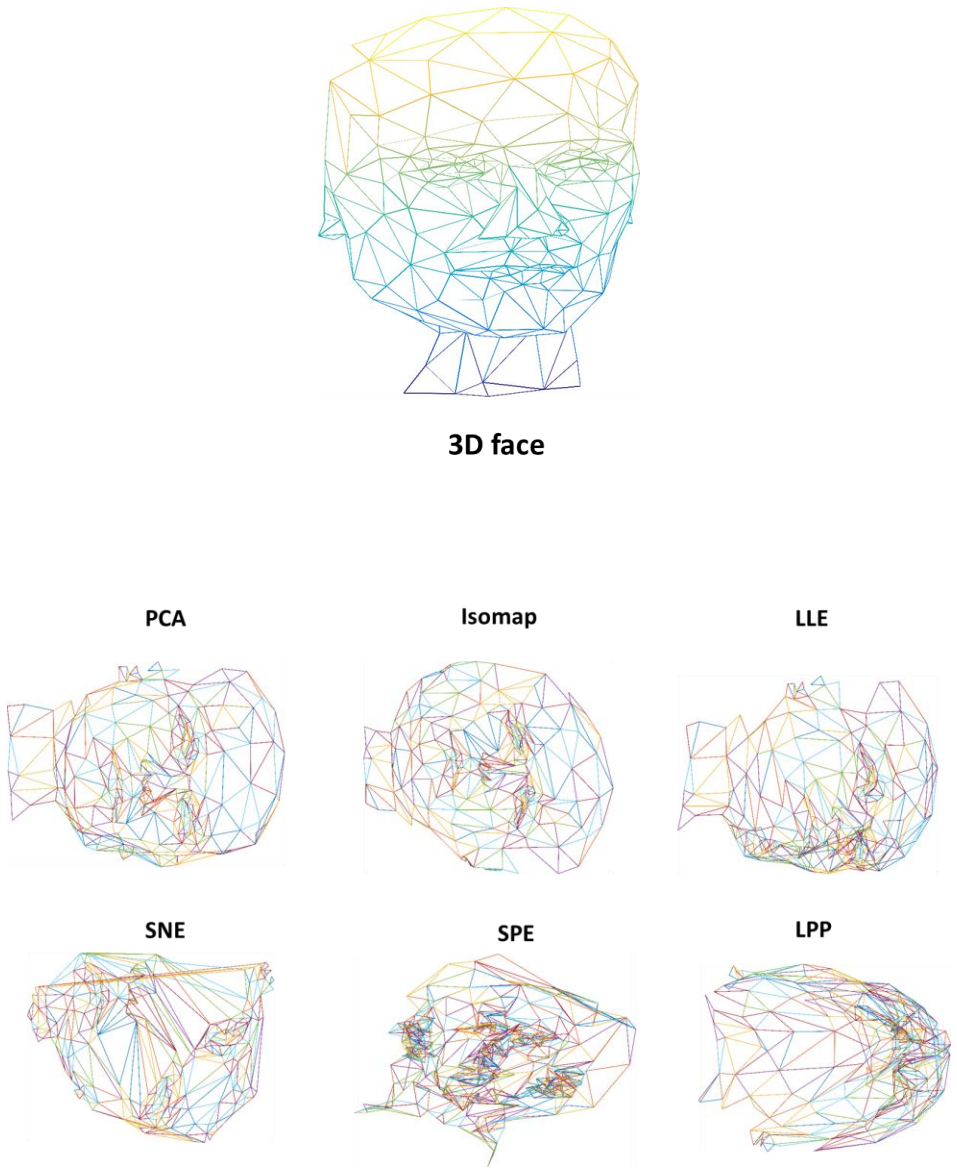

Figure 12 Different dimension reduction results.

Iterative reduction

When dealing with enclosed structures, Isomap is not able to unfold the structures properly. We have proposed another iterative algorithm to unwrap 3-D structures. It removes the fewest edges to open up the structure and unfold it from the opening.

The algorithm starts from placing landmark points on a 2-D plane. Landmark points can be a user defined point set or if not given, vertices connecting to the center point. Any dimension reduction method can be used in here to place landmark points on the 2-D plane. The next step is an iteration that selects the next closest point to the landmark points, and place it on the 2-D plane based on relative positions to the landmark points. To choose the next closest point, the algorithm maintains a distance matrix calculated from shortest path algorithm. After a point is successfully placed, it is set to false and will not be selected again during iteration. There are several cases to consider in placing a point on 2-D. First, the point can be placed inside a triangle, in this situation, the algorithm removes the landmark points one by one and recalculate the new position for that point. If only three landmark points are left and the new position is still not qualified, the algorithm abandon this point to open up the structure. Second, when a new point is placed, its edges can intersect with other edges on the plane, in this situation, re-triangulation is performed to avoid edge conflict.

For algorithm evaluation, hemisphere and sphere are tested in iterative algorithm and Isomap. Hemisphere contains 77 points and 130 triangles. From Figure 13 and Figure 14, it is clear to see both Isomap and iterative algorithm is able to unfold an open structure with preserved triangle size and neighbor relations. However, for an enclosed sphere in Figure 15 and Figure 16, Isomap is not able to unfold the whole structure, on the 2-D triangulation, one side overlaps on the other side, so there is only one side of the sphere is seen, and the other side is hidden below. Iterative algorithm performs better than Isomap, it tries to cut the structure open and unfold it. In the 3-D structure in Figure 15, one can see the opening on the right side of the sphere, which is caused by iterative algorithm.

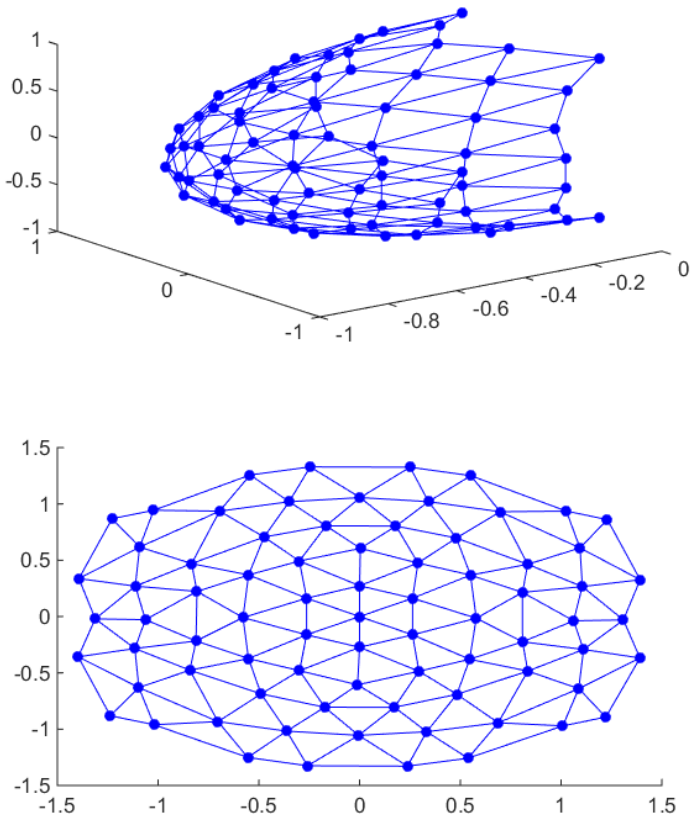

Figure 13 Iterative algorithm on hemisphere. Upper image is the 3-D triangulation of a hemisphere, the bottom image is the 2-D triangulation of the same hemisphere from iterative algorithm.

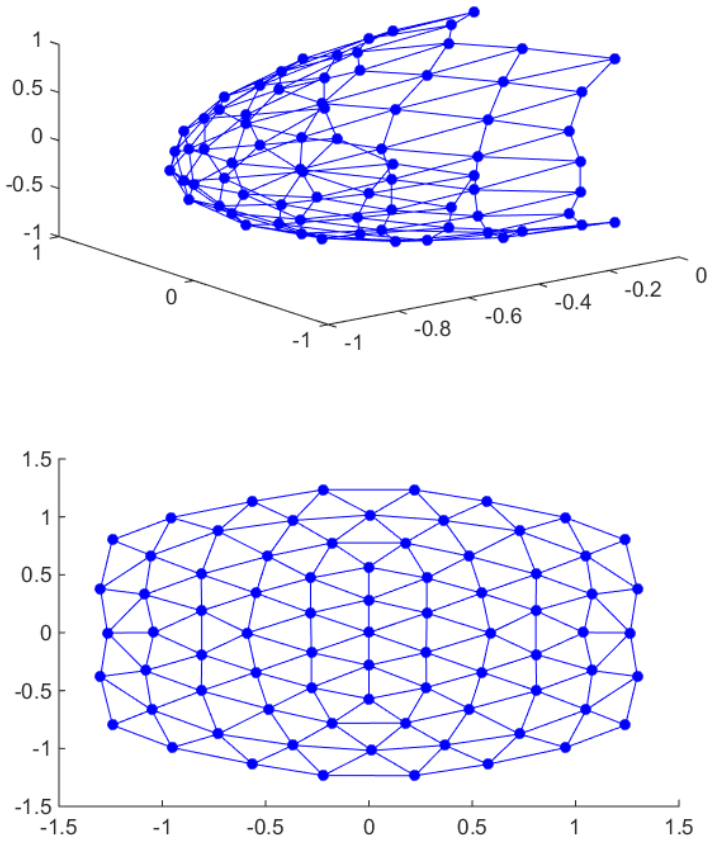

Figure 14 Isomap on hemisphere. Upper image is the 3-D triangulation of a hemisphere, the bottom image is the 2-D triangulation of the same hemisphere from Isomap.

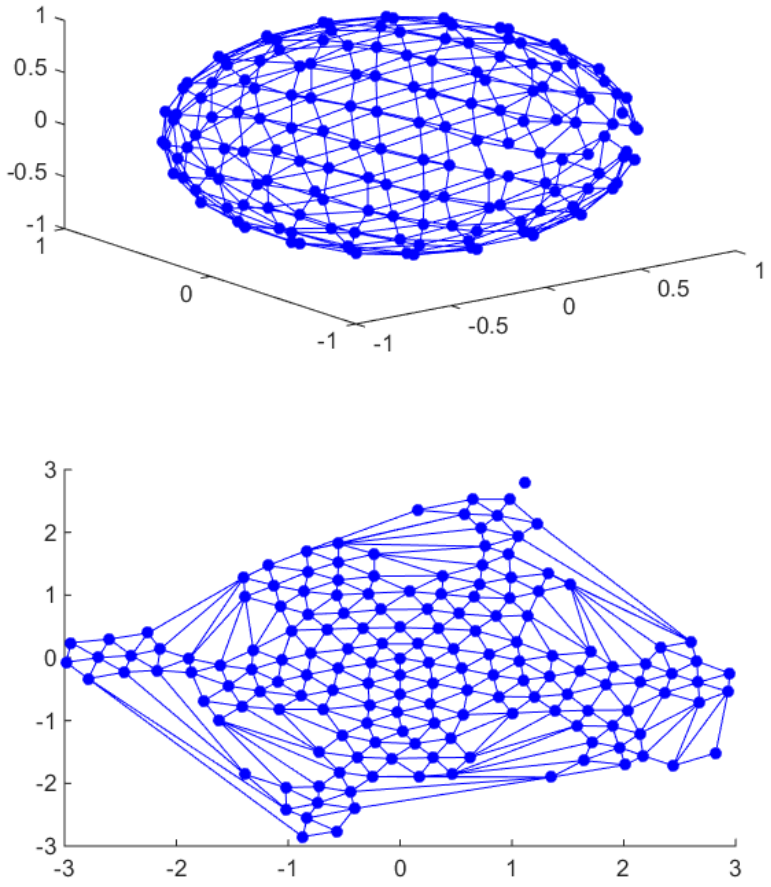

Figure 15 Iterative algorithm on sphere. Upper image is the 3-D triangulation of a sphere, the bottom image is the 2-D triangulation of the same sphere from iterative algorithm.

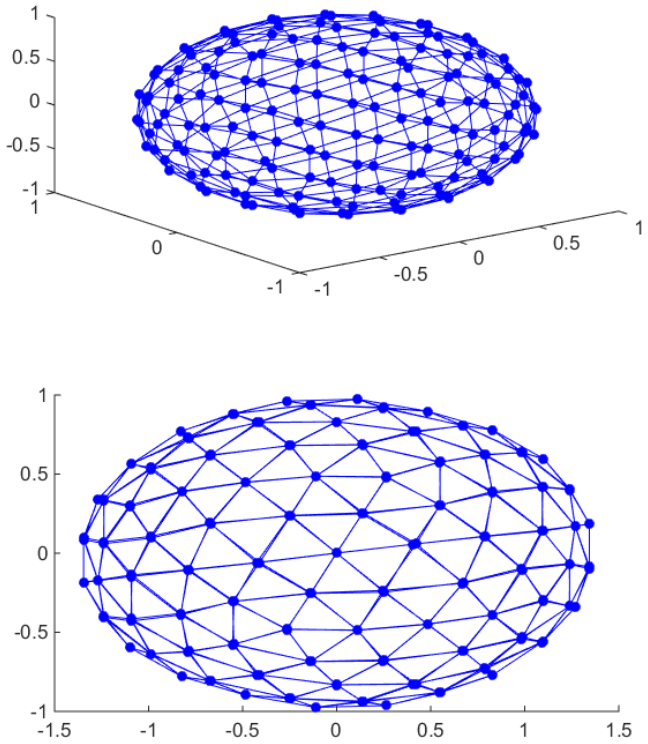

Figure 16 Isomap on sphere. Upper image is the 3-D triangulation of a sphere, the bottom image is the 2-D triangulation of the same sphere from Isomap.

Chapter 4: Template matching

Template matching is a technique in image processing that compares portions of one image to another for object classification. Based on the diverse applications, template matching can be mainly categorized into area-based and feature-based approaches. Feature-based approach [50] is competent in detecting similarities when query image has strong features, such as image corner, edge and other structures that localized in the image. However, feature-based methods are beyond the scope of this study, we will focus on the area-based methods that applicable to the project. Area-based methods are best used in images that do not have apparent features but contain certain differences of pixel intensities. The classical similarity metric are normalized cross correlation (NCC) and square difference. When the template slides on the target image, it records the largest NCC or the smallest square difference for the most similar location. Other metrics are correlation coefficient and Mutual Information. A set of variation of the area-based algorithms are proposed, such as Fast Fourier transformation and the sequential similarity detection algorithms (SSDAs) to enhance the computational efficiency and determine similarity in a far more efficient manner.

We use OpenCV [51] template matching for binding site detection in our algorithm. OpenCV provides various matching metrics, such as square root difference, cross correlation, coefficient correlation and etc., which enables us to select one with best performance. Comparing methods provided by OpenCV are:

CV_TM_SQDIFF:

$$R(X,Y) = \sum_{x',y'} \left(T\left(x',y'\right) - I\left(x+x',y+y'\right) \right)^2$$

CV_TM_SQDIFF_NORMED:

$$R(X,Y) = \frac{\sum_{x',y'} \left(T\left(x',y'\right) - I\left(x+x',y+y'\right) \right)^2}{\sqrt{\sum_{x',y'} T\left(x',y'\right)^2 \cdot \sum_{x',y'} I\left(x+x',y+y'\right)^2}}$$

CV_TM_CCORR:

$$R(X,Y) = \sum_{x',y'} \left(T\left(x',y'\right) \cdot I\left(x+x',y+y'\right) \right)^2$$

CV_TM_CCORR_NORMED:

$$R(X,Y) = \frac{\sum_{x',y'} \left(T\left(x',y'\right) \cdot I\left(x+x',y+y'\right) \right)^2}{\sqrt{\sum_{x',y'} T\left(x',y'\right)^2 \cdot \sum_{x',y'} I\left(x+x',y+y'\right)^2}}$$

CV_TM_CCOEFF:

$$R(X,Y) = \sum_{x',y'} \left(T'\left(x',y'\right) - I\left(x+x',y+y'\right) \right)^2$$

CV_TM_CCOEFF_NORMED:

$$R(X,Y) = \frac{\sum_{x',y'} \left(T'\left(x',y'\right) \cdot I'\left(x+x',y+y'\right) \right)^2}{\sqrt{\sum_{x',y'} T'\left(x',y'\right)^2 \cdot \sum_{x',y'} I'\left(x+x',y+y'\right)^2}}$$

Where: $T'(x',y') = T(x',y') - \frac{1}{w \cdot h} \cdot \sum_{x'',y''} \left(T\left(x'',y''\right) \right)$

$I'(X+X',Y+Y') = I(x+x',y+y') - \frac{1}{w \cdot h} \cdot \sum_{x',y'} \left(I\left(x+x'',y+y''\right) \right)$

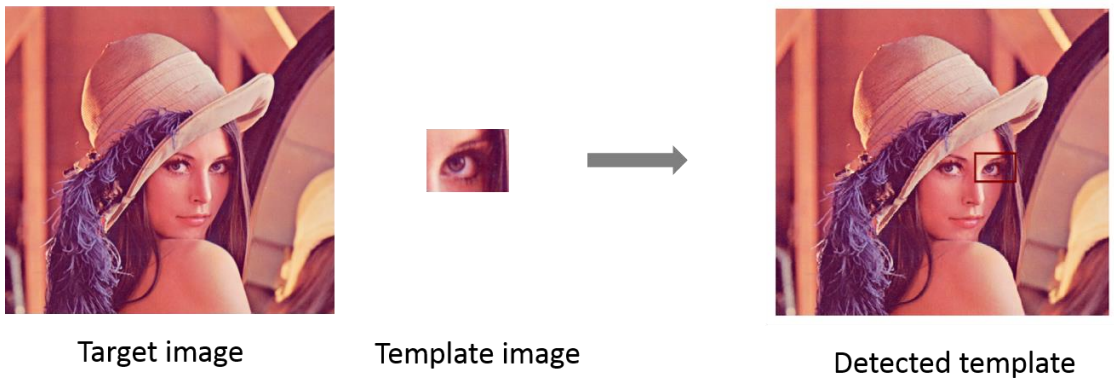

Figure 17 Template matching example.

Template matching is used in our application to detect the similar areas on a pair of protein surfaces generated from dimension reduction. Since protein structures have different orientations, and when it is mapped on 2-D dimension, the rotation of the image is changed. Template matching is unable to detect rotation. We have proposed an algorithm that tackles the rotation problem. The algorithm generates an expanded outer bounding box that wraps the template window inside, and the template window can be rotated inside the bounding box. The rotated template window is used for template matching. The details of this algorithm is elaborated in 0.

Here is one example of template matching detection. Target image is a face image, template image is cropped from the target image. The template matching is used to detect the correct location that matches the best on target image. Figure 17 shows the process of a template matching, with input of target image and template image, the result is the detected location on target image which matches the most with template image.

Chapter 5: Texture mapping

Application workflow – texture mapping

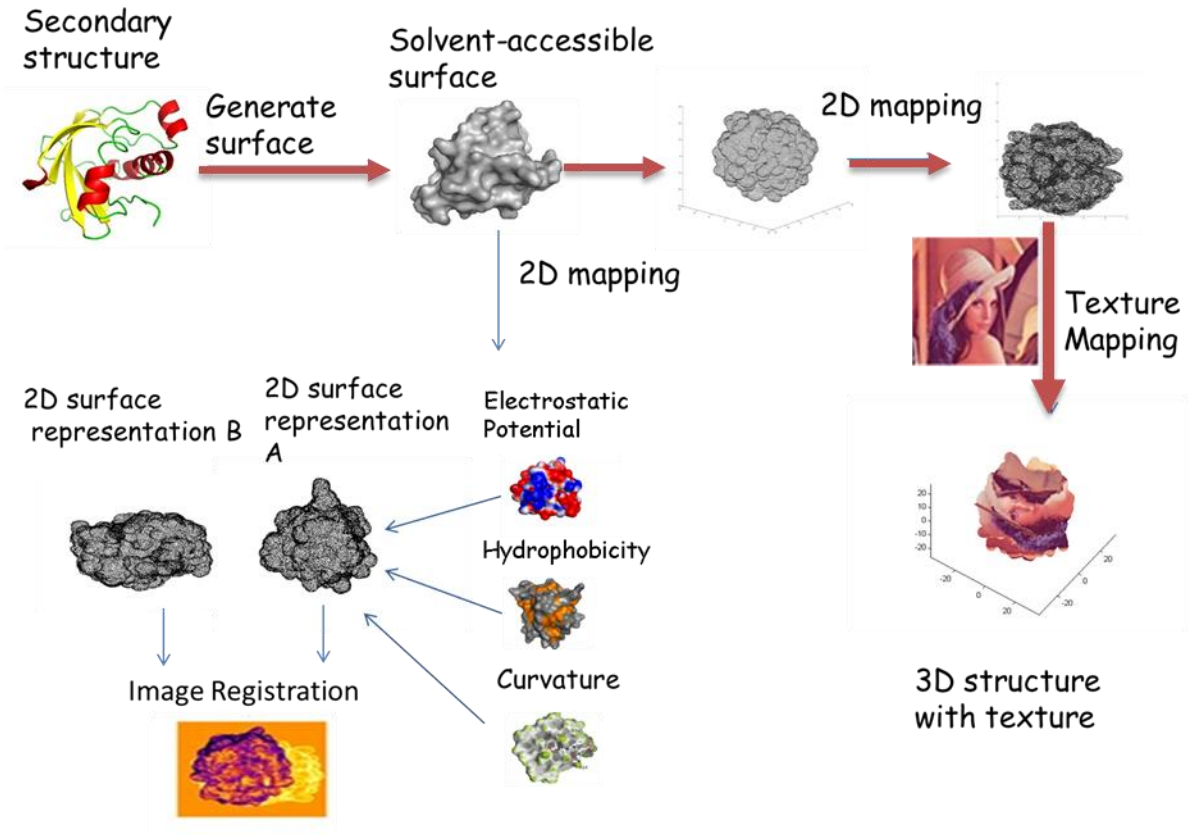

Figure 18 Texture mapping Application workflow.

There are two major workflows in this application. The first direction is texture mapping. The large brown arrow in Figure 18 is the direction of texture mapping. A texture image can be superimposed onto a 2-D surface representation, and points on the surface are assigned the pixel values based on the location. Those points with pixel values are projected back on 3-D points so that the texture image is mapped on the 3-D structure.

The second direction is the surface comparison. Dimension reduction maps 3-D surface to 2-D, with physicochemical and geometrical features assigned on the surface. Template matching is used to maximally align the two surfaces (images) to find the local similar region.

Texture mapping

Texture mapping is a technique that applies a texture image on a 3-D object to enhance visual richness in computer graphics and image processing. It provides surface details to a plain 3-D object by modifying surface color, surface normal, and transparency and so on. Previous literatures have explored plenty of texture mapping algorithms, and commonly, there are two ways to implement it, one is through adding more polygons to render the surface details, and the other is mapping a texture image to the surface. We will be only discussing the last method in here.

In our algorithm, a 3-D object is triangulated to form a triangular mesh with vertices and edges. We have demonstrated that a 3-D object can be reduced to 2-D by using dimension reduction. Since we are not trying to unfold the structure mentioned before, having triangles overlap is allowed in here. Principle component analysis (PCA) and Isomap are used for reducing dimensionality. The 2-D points are superimposed onto a texture image and the each point is given the pixel value of the texture image based on its location on the image. Figure 19 shows an example of the texture mapping procedure. The texture image is a black and white chess board, and the 3-D points is a random point cloud, shown in the first image of second row. The 2-D points are mapped from 3-D points, and superimposed on the texture image. Points locates on black square are assigned with black color, and the rest locates on the white square are assigned with

white color. The middle image of the first row shows the 2-D points after color assignment. When the color interpolation is computed on the 2-D points based on the triangular patch of the mesh, we are able to recover the check board on the 2-D points, shown in the right image at the first row in Figure 19. The corresponding 3-D points are shown in the first image of the second row, and the color value of 2-D points are assigned back to those 3-D points, with color interpolation, the 3-D object has a chess board texture on its surface, shown in the right image of the second row.

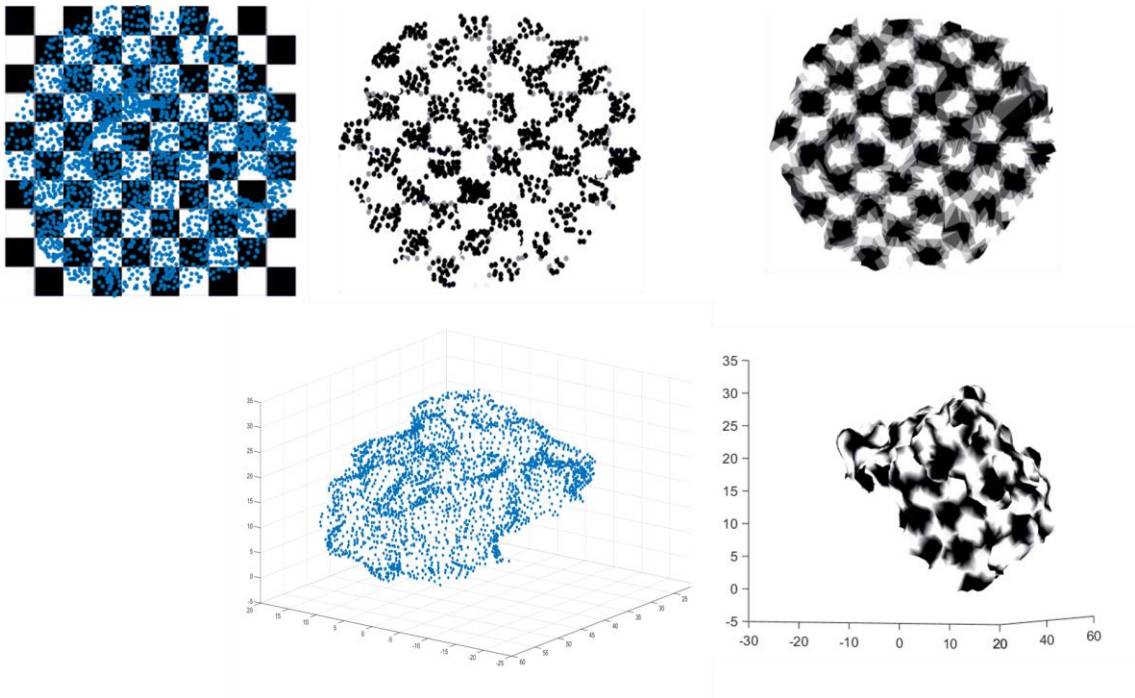

Figure 19 Texture mapping example with the chess board image.

Texture mapping results

Figure 19 is used to explain the texture mapping procedure on the previous section, but it is also a good texture mapping example. The 3-D object in Figure 19 is the binding site section of protein 1mh1, containing 2953 points. The corresponding 2-D points are computed using Isomap, and are superimposed on a chess board image, shown in the first image of first row. From the right figures of the first and second row, it is clear to see the chess board properties are well maintained on both 2-D and 3-D object.

Figure 20 and Figure 21 are another two texture mapping examples. In Figure 20, the 3-D hemisphere is applied with an earth image. The first image is the 2-D points with mapped texture image after color interpolation, and the last image is the 3-D object mapped on the earth image. It shows the texture image are well captured in 3-D object. Figure 21 uses protein 1crn as the 3-D object and a floral image as the texture image. The first image is the 2-D points overlapped with the floral image, and the right image is the 3-D object with floral texture image mapped on the surface. The layers of color and strip pattern on the 2-D image are well preserved on the 3-D object.

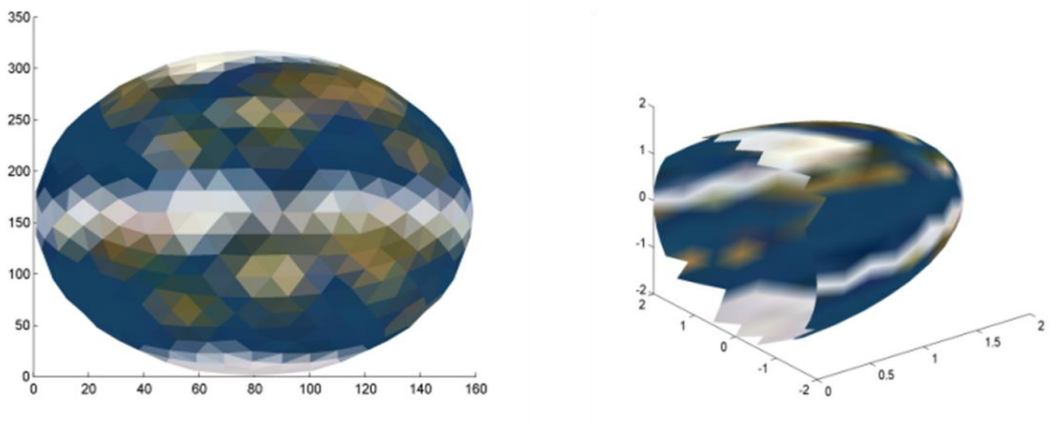

Figure 20 Left: overlap of 2-D sphere surface and template image. Right: 3-D texture mapping.

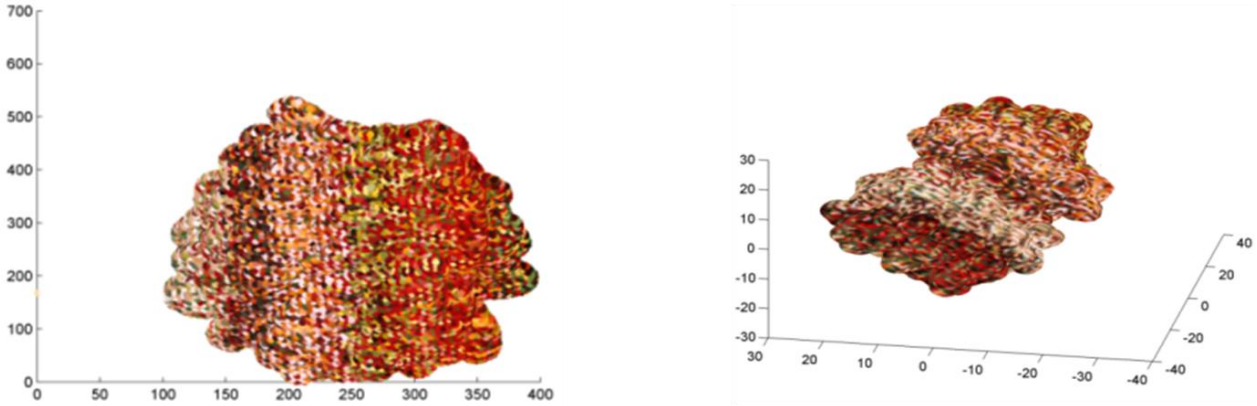

Figure 21 Left: overlap of 2-D protein surface and template image. Right: 3-D texture mapping.

Chapter 6: Protein surface Comparison

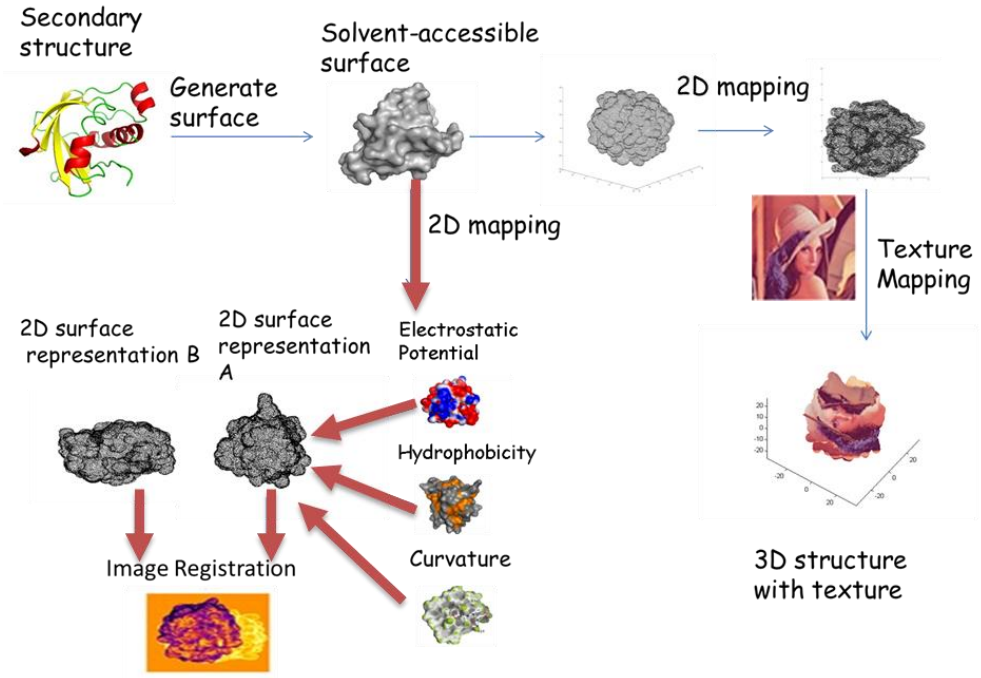

Figure 22 Application workflow of surface comparison.

The second direction of the workflow is shown in big arrows in Figure 22. We have proposed a method that utilizes both geometrical and physicochemical features for molecular recognition. Figure 23 is another explanation of the workflow. The workflow starts from the surface generation of a given protein, followed by feature enrichment of physicochemical and geometrical information. This results in the color attributes on the 3-D surface. Protein segmentation is used to open up the enclosed 3-D protein structure, which is a crucial step for dimension reduction in that dimension reduction is unable to perform well on an enclosed structure. Segmentation can be done by general cut or binding site cut based on the knowledge of binding site information. Dimension reduction

is utilized after to unfold the structure, resulting corresponding 2-D images. Template matching is the final but key step to quantify the similarities between two proteins.

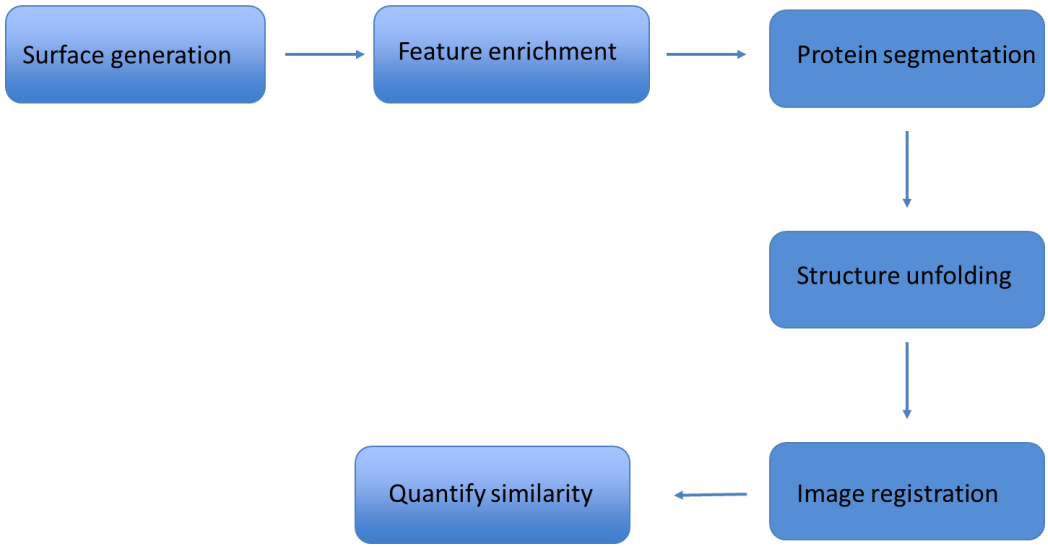

Figure 23 Another illustration of protein surface comparison workflow.

Enrichment with surface features

Geometric features of the protein surface are important functional determinants and have been used by some methods as the sole source of information for comparing protein surface regions. However, it is well-known that biochemical properties play an important role in determining binding interactions and enzymatic activity [52]. These biochemical properties are often captured in the form of amino acid or atom types when a residue-wise or atom-wise matching is performed between surfaces being compared [53].

Without loss of generality, we consider the following properties for each surface point: hydrophobicity, electrostatic potential, curvature, and evolutionary conservation. These

properties are commonly used in studying protein folding, protein-protein interactions, protein-ligand binding, and enzymatic activity.

Hydrophobicity property has been one of the most used properties in studying protein structure and folding [54] . We calculate the hydrophobicity values at each surface point using VASCo, which takes advantage of distance-dependent contributions of molecular lipophilicity potential from individual atoms [55].

Electrostatic potential plays a central role in molecular interactions, for example, clusters of charged and polar residues enhance stability of protein-protein complexes [56] and electrostatic potential guides ligand-binding interactions [36]. We use the DelPhi program for calculation of the electrostatic potentials at each surface point [57].

Curvature is an intrinsic property capturing local geometry of a surface and has been useful in assessment of protein-ligand binding interactions [58]. For each surface point, we calculate the Gaussian curvature from the surface triangulation [59, 60], where points from the convex and concave surface regions take on positive and negative curvature values, respectively.

The final property for each surface point is evolutionary sequence conservation of the surface residues. Protein binding sites usually display a high level of conservation, and it provides an indirect indication of catalytic and ligand-binding activity of the surface residues, because these residues are under greater evolutionary pressures [36]. The calculation of the conservation weights is based on the observation that the importance of a residue is reflected in its evolutionary conservation; the more important a residue is, the sooner it becomes fixed in different evolutionary branches. Important residues are more likely to result in a loss of the protein functions if they mutate into other residues. Thus,

we quantitatively predict the relative importance of the residues in a protein by calculating the entropy of each position and assign larger weights to those more conserved positions. To calculate evolutionary conservation, we first perform a PSI-BLAST search of the protein sequence against NCBI non-redundant protein sequence database [61] and multiply align the search results using MUSCLE [41]. We then derive sequence conservation scores for each residue using the method described in STACCATO [42]. Each surface point is assigned the conservation score of the residue it is closest to.

Since image processing techniques handle best on the traditional RGB encoded images, we enrich the protein with three properties at a time, results in four different groups of color combination, shown in Table 2. Each group is represented by its color code, which consists of the initial letters of three features respectively.

Table 2 Four groups of color combination

color code	Red	Green	Blue
EHV	electrostatic potential	hydrophobicity	curvature
CHV	conservation	hydrophobicity	curvature
ECV	electrostatic potential	conservation	curvature
EHC	electrostatic potential	hydrophobicity	conservation

When the protein surface is mapped to a 2-D surface, the features associated with each point are also carried over. We then convert the 2-D mapping to an image where each pixel takes on the average values of the features for the points that map into that pixel location. When an active site or another surface region of interest is defined, the points in the image corresponding to that site are used to generate a minimum bounding box enclosing all such points. Although the active site can be more precisely defined by a polygon mask, for computational simplicity we represent the active site region using the smallest rectangle that encloses these points.

2-D image generation

We have discussed about the method to calculate binding site points for a 3-D structure. The problem here is how to derive the pixel location of a binding site region from a 2-D image. Since dimension reduction only returns the 2-D coordinates of original 3-D vertices. We need a method to derive the relation between point coordinates and pixel locations. The way we solve it is to draw the 2-D points on an image and measure the image's length and height, and the number of pixels for each unit in length and height. Based on the densities on both height and length and the coordinates of binding site points, we are able to calculate the pixel location for each point. After the relation is obtained, the binding site region can be mapped on the 2-D image. Since binding site vary in shapes and size, the actual binding site regions are irregular polygons. We define a minimum outer bounding box to include the complete binding site region for each protein, so that in template matching algorithm, the inner part of the bounding box can be rotated to compare against the target image. Figure 24 gives an overall procedure from

template window generation to template matching. The first image is the 2-D point cloud with color enrichment. The second is the interpolated image, color interpolation fills up the empty space among points. With binding site surface points highlighted in blue dots. In the third image, a bounding box (rectangle) is generated based on binding site points, and the template window is extracted from the surface and used to compare with target image, as shown on the last image.

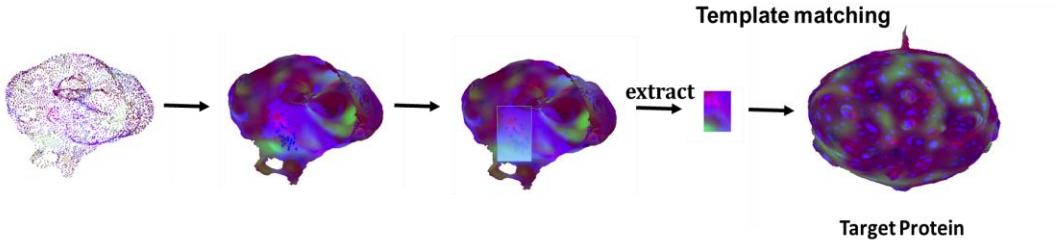

Figure 24 Template matching process between template window and target image.

Comparison between two proteins with binding site of one protein is known

The binding site region of one protein is known, and we want to see if we can utilize the known binding site information of one protein to locate the binding site region of another protein. A high similarity score from the template matching indicates two images share a similar sub region. In this application, the known binding site region is used as template image, and slides on the target image of the query protein to quantify similarities. The query protein is cut into multiple sub sections since the protein is unknown, and the corresponding 2-D images are created for each section respectively. A collections of 2-D images is compared with the template image respectively, and template matching returns

a similarity score rank and a corresponding location on those 2-D image respectively. The image with the highest score and its corresponding location on target image predicts the binding site region for the query protein.

Binding site region rotation

In order to maximally match the binding site region from different orientations, we incorporate the image rotation into template matching. Our algorithm rotates the template window every 10 degrees, and each newly generated window compares with target image. Thus, template matching is able to capture relevant information even if proteins are in different 3-D orientations. Figure 25 shows the rotation process. The left upper image is a rotated template image with template window highlighted in red rectangle. The target image is at the button left. Each rotated template window compares against the target image and results in a 3-D scoring matrix. Each element in the scoring matrix represents the similarity score at the location of the comparison, and with the rotation angle of the template image at that location. The middle and last image in Figure 25 shows the representation of the 3-D matrix with x and y values indicate the location of the template window, and z value indicates the rotation angle.

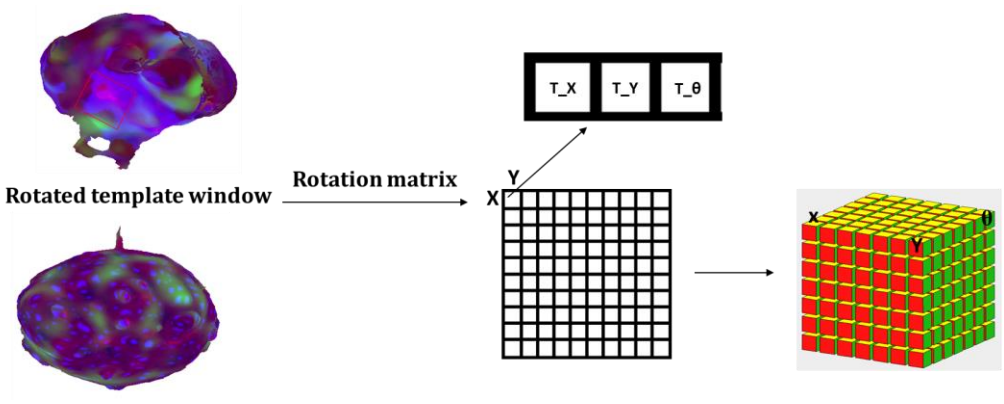

Figure 25 Template matching with template window rotation.

Results

For each surface point, the closest amino acid residue is calculated by shortest path algorithm. Surface points are set to be binding site points if corresponding residues locate on binding sites according to Catalytic site atlas [62]. When the binding site residues are not provided, the closest surface points to the binding ligand are calculated and assigned as binding site points.

ALDH superfamily

The first case study is ALDH superfamily. ALDH enzymes plays a crucial role in aldehyde detoxification by catalyzing aldehydes to carboxylic acids. Its active sites have been highly conserved over time, and share a number of conserved residues for catalysis, such as Cys-302, Glu-268 and Asn-169 [36]. Two proteins are chosen from the superfamily: rat liver ALDH3 (pdb: 1ad3) and sheep ALDH1 (pdb: 1bxs). Sequence analysis using BLAST shows only 29% sequence identity between them, but their binding sites of ligand NAD are hugely conserved. The goal of this study is to evaluate the accuracy of template matching. The binding site section for each protein is segmented

and the 2-D binding site images are computed respectively. In this study, we use color code EHC (electrostatic potential, hydrophobicity, evolutionary information) for template matching. The template window of the binding site image for protein 1ad3 is generated and slides left to right and up to bottom on the binding site image (target image) for protein 1bxs to obtain the best matching location. The best matching location is defined as the location where the minimum square root difference of pixel values between template window and target image occurs.

In Figure 26, the color features are maximally preserved on 2-D images after dimension reduction. The row illustrates the transition from 3-D binding site section to corresponding 2-D image for each protein. Red polygon on 3-D sections are the binding site regions for each protein, and the red rectangle on 2-D image is the template window generated based on binding site region. The template window slides on the target image (bottom right), and the predicted binding site region on the target image is drawn with green rectangle. The actual binding site region for the target image is drawn with red rectangle. The green and red rectangles on the target image have a large overlap from one another, which demonstrates that given two binding site images, template matching is able to detect the binding site region based on the color information. The binding sites on both images have the strong purple area, which is a mixture of the red and blue color channels, which illustrates the importance of electrostatic potential and evolution conservation for these two proteins.

We also tested on color code EHV for previous two proteins. Figure 27 shows the actual and predicted binding site region for the same protein pair. The color feature of color code EHV does not strongly distinguish the binding site from neighbor areas compared to

color code EHC in Figure 26. However, the binding site region still shows a strong conserved electrostatic potential (red color). In this case, color code EHV is unable to detect the actual binding site well in that the predicted location does not cover the correct binding site region. We think this is due to the strong conservation on the binding sites for both proteins that lead to the correct prediction for color code EHC, and when the important information is lost in color code EHV, the prediction is not able to find the similar region.

We have shown the 3-D surfaces of the two proteins on both color code EHV and EHC in Figure 28 and Figure 29 respectively. The overall 3-D structures are different for two proteins, however, the binding site region (pockets in the middle) have similar shapes. The white balls inside the pockets are the ligands for each protein. In Figure 28, the binding site region do not have strong similarity stick out, and the peripheral area is also not showing the strong correlation. However, in Figure 29, when the evolution conservation is added in the color code, the binding site region are both covered with strong purple pattern, especially inside the pockets. This similarity provides a strong information for template matching to find the correct location.

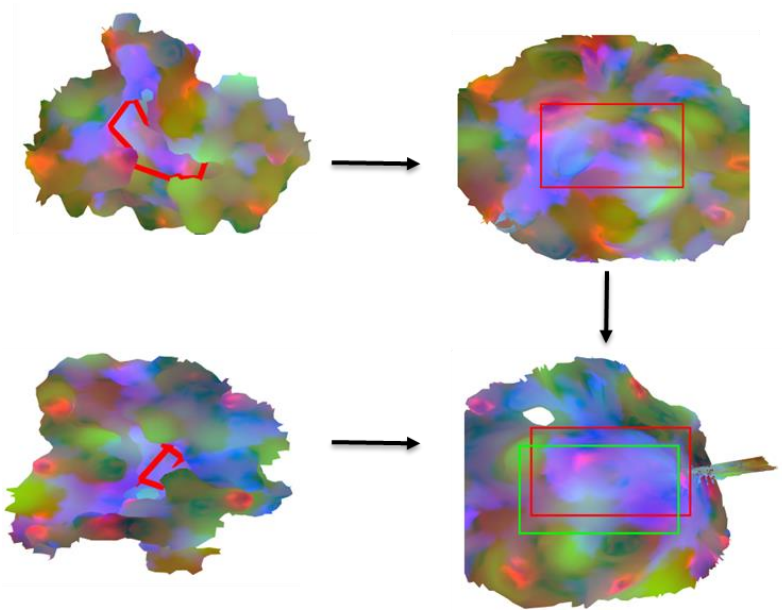

Figure 26 Template matching between protein 1ad3 and 1bxs. Left: 3-D structure of protein 1ad3 and 1bxs, with their binding site region plotted with red polygon. Right: corresponding 2-D images. Red rectangle is the correct binding site region and green rectangle is the predicted binding site region.

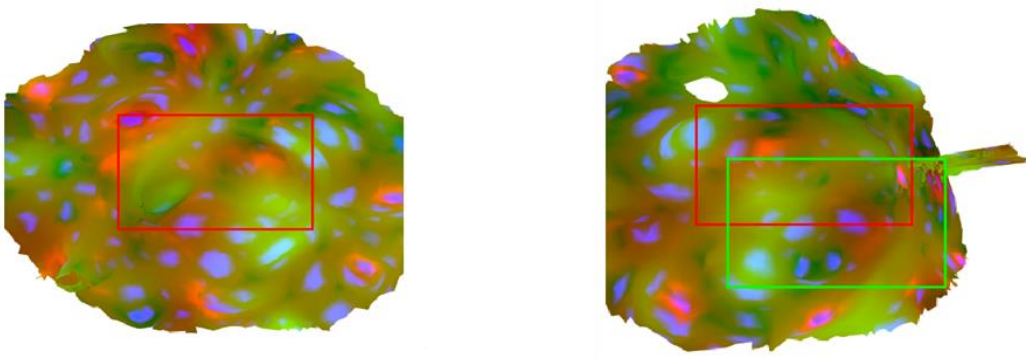

Figure 27 Template matching between protein 1ad3 and 1bxs. Left: 2-D image of protein 1ad3, with binding site region plotted with red polygon. Right: 2-D image of protein 1bxs. Red rectangle is the correct binding site region and green rectangle is the predicted binding site region.

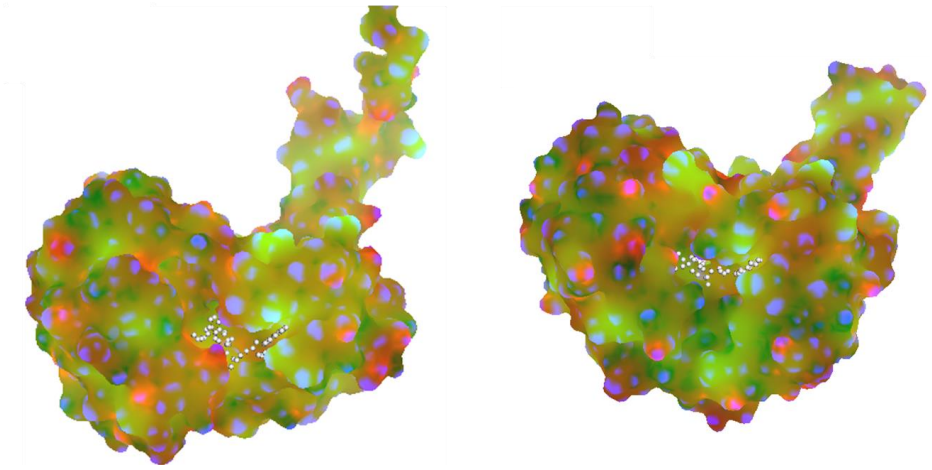

Figure 28. 3-D surface of protein 1ad3 and 1bxs with color code EHV. White balls are ligand NAD.

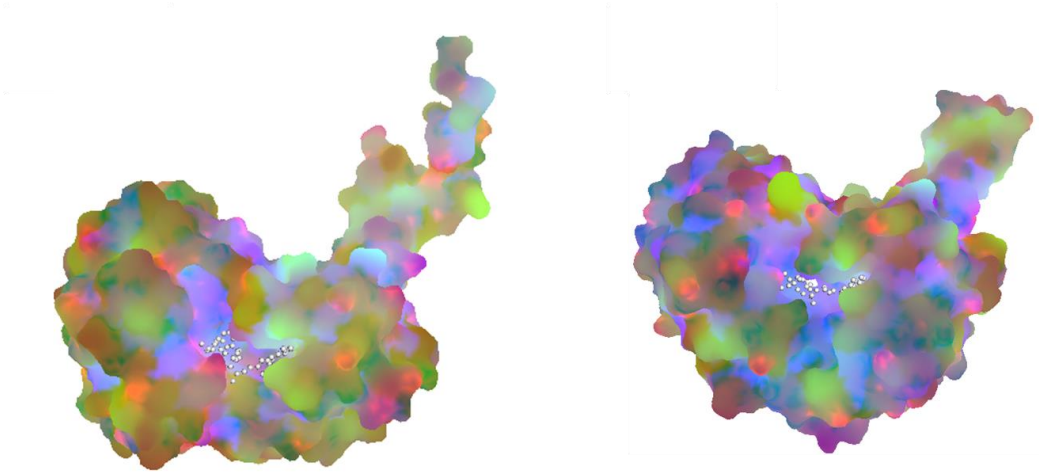

Figure 29. 3-D surface of protein 1ad3 and 1bxs with color code EHC. The purple color shows the strong evolution conservation in the binding site areas.

Serine proteinases

The second protein pair is two serine proteinases. In general, serine proteinase has three major amino acid residues called catalytic triad. The three residues can be far apart in the primal sequences but in tertiary structure, they tend to form a certain pattern that performs catalytic action. Mutations may leave the binding pocket unchanged but to alter the rest of the structures. Two serine proteases are chosen from protein data bank [63]: (pdb: 1trn) and (pdb: 2ptn). Although these two proteins only share 38% sequence similarity according to BLAST, their binding sites are formed with equivalent major amino acids: serine, histidine and aspartate. In this study, the binding site region of protein 1trn is calculated based on the position of catalytic triads and its binding site image and template window is determined. We assume the binding site of protein 2ptn is unknown, and it is segmented using general segmentation. The general segmentation results in different number of protein sections depending on the size of a protein. It limits the radius of each section to at most 15 angstrom, and it also defines the exclusion radius to prohibit points within it from being re-selected. The default value of exclusion radius is 10 angstrom. Points within exclusion radius to the center of the patch in one section will be marked as inactive, however, points beyond exclusion radius can be re-selected by other sections to guarantee enough section overlap. These two radiuses ensure that there are enough sub sections for computation but not too many to cause computational burden. In this case, protein 2ptn is segmented into 62 sections. The template window of protein 1trn slides on each of 62 images and the similarity score is computed, from which, the binding site of 2ptn is predicted. Electrostatic potential, hydrophobicity and evolutionary information are used to fill RGB channels respectively.

Figure 30 shows the result of the prediction. First row is the procedure of generating binding site image for protein 1trn. Red polygon shows the actual binding site region, and binding site image with template window is obtained, shown in last image. The procedure is from 3-D surface generation, through binding site region determination, binding site section calculation to binding site image generation. The second row is the procedure of a general segmentation for protein 2ptn. One section is shown as an example. The green rectangle on last image is the predicted binding site region and red rectangle is the actual binding site region for protein 2ptn. The second row procedure is easier, which only needs 3-D surface generation, general segmentation computation to obtain sections. Template matching is calculated after the first and second procedure, and the input to the matching is the binding site image and 62 images generated from general segmentation. In

Figure 30, one can see a very unique purple color and shape pattern around binding site regions. This is due to the strong amino acid conservation feature of the three major amino acids.

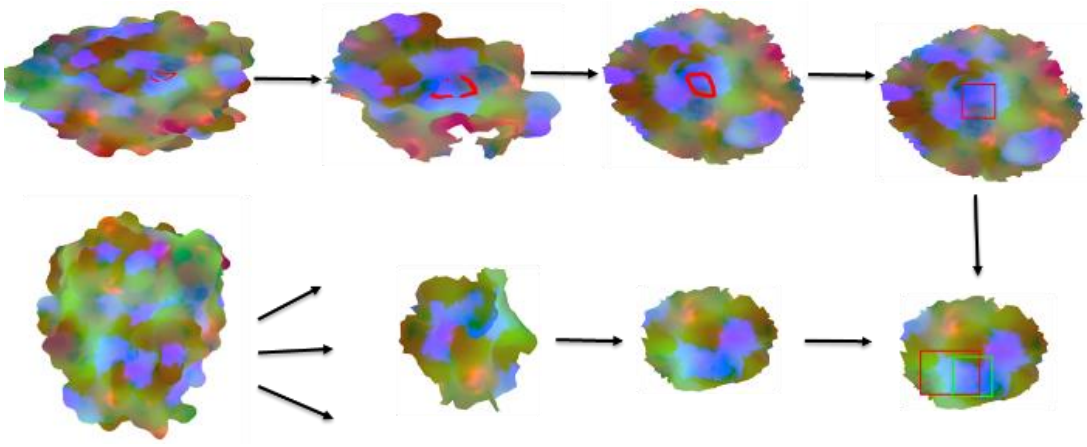

Figure 30 Template matching between protein 1trn and 2ptn. First row is the binding site segmentation for protein 1trn. Second row is general segmentation for protein 2ptn. Red polygon is the binding site region. Green polygon is the predicted binding site region.

Human Rac1 and HRas

The third case study is protein human Rac1 (pdb: 1mh1) and HRas (pdb: 4g3x). Rac1 is a member of Rho family and downstream effector of Ras. HRas is a member of Ras that operates as molecular switch on the inner surface of the plasma membrane. Both Ras and Rho are the sub families of Ras super family. They both share the common GNP binding sites but they function at different targets. Blast shows 31% sequence identity between two proteins.

Four color codes are computed respectively in this study. Assume binding site of protein 1mh1 is known, and is used to detect the unknown binding site for protein 4g3x. The template window is computed based on the binding site region of protein 1mh1, which slides on 53 images of protein 4g3x that are generated using the general segmentation. In Figure 31, the row shows the binding site image of protein 1mh1 and actual and

identified binding site region for protein 4g3x. Each row represents a color code. The red window on the left column is the template window for protein 1mh1. The actual and predicted binding site of protein 4g3x are shown in red and green rectangle respectively. The result shows that by using color code EHV, CHV and ECV, we are able to accurately predict the binding site region. The actual and predicted binding site regions are almost overlapped. However, prediction from color code EHC shifts to the right of the correct binding site region, thus the prediction is not accurate. This examples demonstrates again that the physicochemical properties are crucial in determining the binding site location, and different proteins are sensitive to different properties. We think that protein pair 1mh1 and 4g3x must have similar shape and size in the binding site region so that when curvature is removed from the color code, the prediction becomes inaccurate. In order to test our theory for the inaccurate prediction on color code EHC. We have drawn the 3-D surface of each protein with a clear view on the binding site pockets in Figure 32. The small white balls are points of ligand GNP, and the cavities attached to ligand GNP are the binding site pockets for each protein. It is clear to see despite of the different overall structures, the binding site region have a very similar shapes and sizes: both are long and have an oval pocket shape, which denotes the similar curvature value.

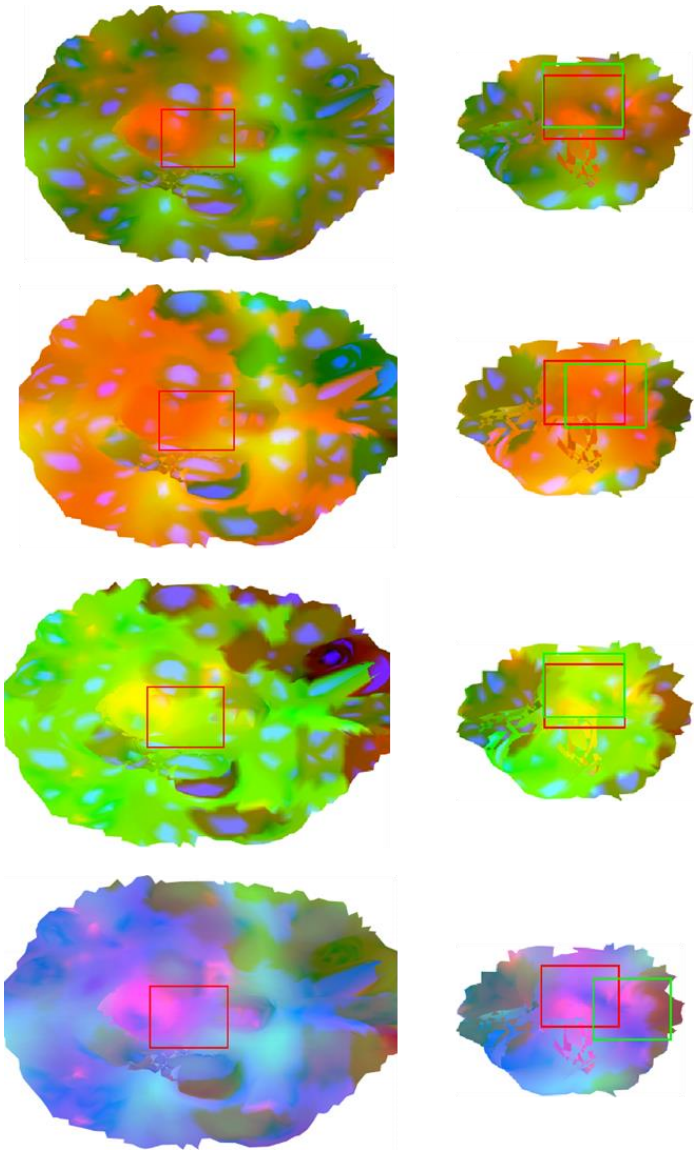

Figure 31 Template matching of protein 1mh1 and 4g3x. Left: binding site images of protein 1mh1 from different color combination, red window locates the correct binding site region. Right: the detected binding site images of protein 4g3x with the corresponding correct (red) and predicted (green) binding site region.

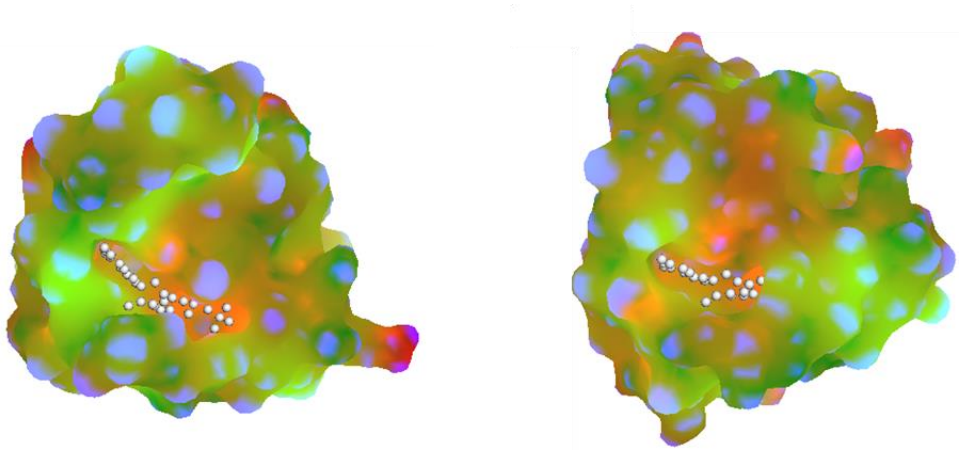

Figure 32 3-D structures of protein 1mh1 and 4g3x. The long pockets are binding site locations, and white balls are ligand GNP.

Proteins with different SCOP classification

We have also selected 8 proteins from PDB, the selection is just using keyword search. The 8 proteins bind with four different ligands, and we have purposely selected protein pairs that are in different SCOP levels. Each pair of proteins diverge at either top 1 or top 2 SCOP level. The higher the SCOP lever is, the more divergent of proteins are. Table 3 shows the protein id, SCOP classification, and binding ligand. The first column is protein pairs with pdb id, the second column is their differences at the SCOP level, and the third is the ligand name. The binding site patch for protein 1jh8,2biu,2hgs,2abj are computed, their counterparts in the same ligand group, which are 1jys, 1q1c,1dug,1h0c, are segmented into multiple sections, results in 60 sections for protein 1jys, 82 sections for 1q1c, 74 sections for 1dub, and 105 sections for 1h0c. Binding site section is then used to detect similarity on their counterpart proteins. We have tested on four different color codes, and chosen the best color code prediction for each pair.

Table 3 SCOP classification

Proteins (pdb)	Different at level	Ligand
1jh8-1jys	Fold	Adenine
2biu -1q1c	Class	Dimethyl sulfoxide
2hgs-1dug	Fold	Glutathione
2abj-1h0c	Class	Pyridoxal Phosphate

The predicted binding sites are shown in Figure 33. Protein pairs for each row are known proteins to unknown proteins. First column is the binding site section for known proteins, and second column is the predicted sections and its predicted binding site location for unknown proteins. The red rectangle represents the actual binding site region, and the green rectangle is the predicted binding site region. All the actual binding sites overlap with the predicted binding site regions, and the color features for actual binding site regions of each pair have similar patterns. The tertiary structures of each protein pair is shown in Figure 34. The figure shows even though proteins are in different folding, different tertiary structures, our algorithm is still able to using the binding site information of one protein to successfully predict the binding site of another protein.

We have selected protein 2abj and protein 1h0c, as an example to show their similar 3-D binding site patterns in Figure 35. It is clear to see even though their overall structures are different, the binding site patterns are very similar, and both show high electrostatic potentials and evolution conservation around the binding site region. The white ball

inside pockets are ligand PLP points. The pocket size, shape and color are similar with each other, even the peripheral region around the pocket share similar pattern.

The example demonstrates our algorithm is able to detect the local binding site in the absence of global similarity. It is not only able to identify the correct section where the binding site resides among a large number of segmentations of an unknown proteins, but also correctly predict the actual binding site region.

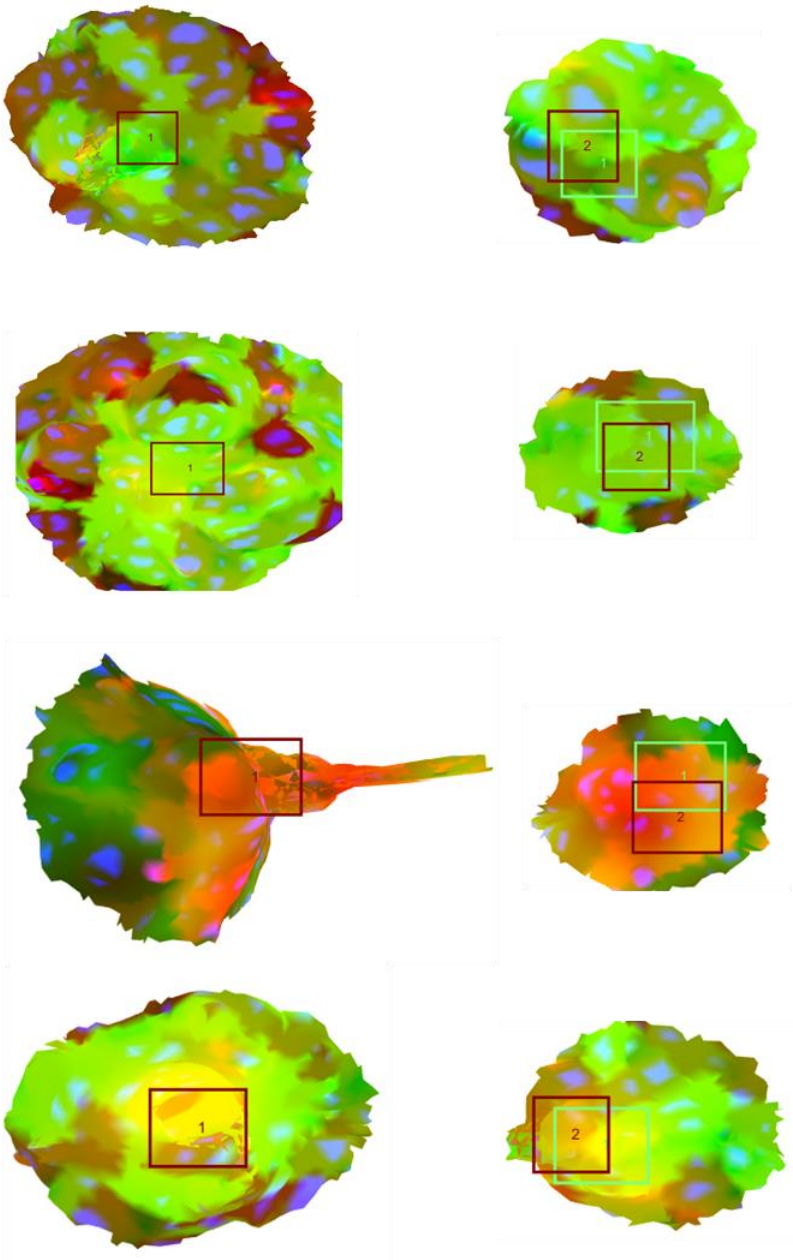

Figure 33 Prediction results for each protein pairs. First column is the binding site section for known proteins, and red rectangle is binding site region. Second column is the predicted section and binding site location for unknown proteins. Red rectangle is actual binding site and green rectangle is predicted binding site.

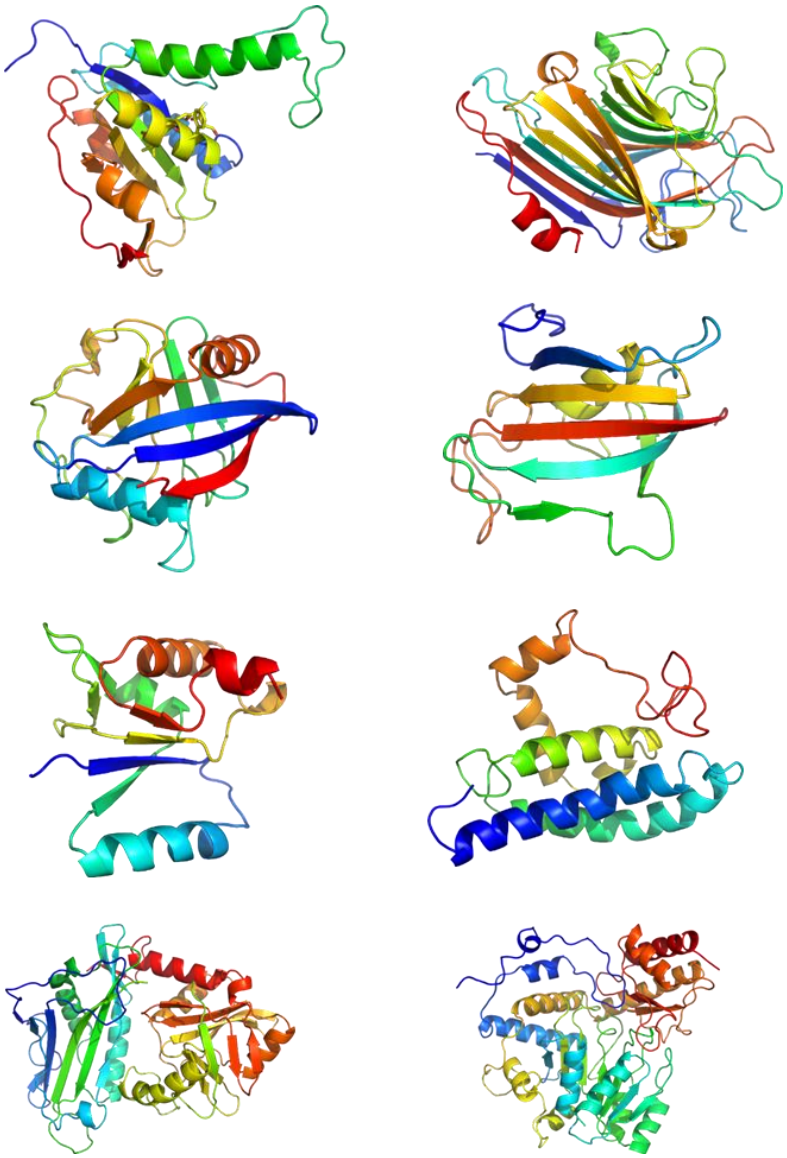

Figure 34 Tertiary structures of protein pairs.

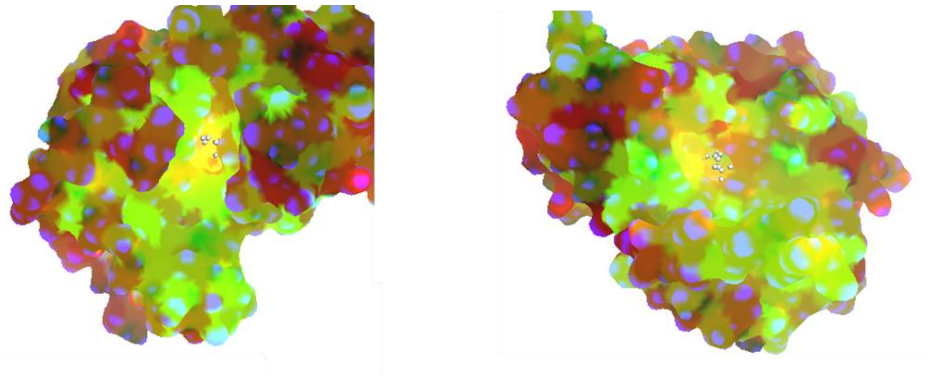

Figure 35 Example of protein pair 2abj-1h0c. Left image is the 3-D structure of 1h0c, and right image is the 3-D structure of protein 2abj. The white balls inside pockets are ligand PLP.

Chapter 7: Drug target prediction

Historically, the discovery of drugs involve with experimental active ingredient identification and by serendipitous discovery. However, with the advent of the massive protein discovery and computational technologies, computer-aided drug design has been brought up and enjoyed much of attention. The drug design technique embraces the design of compounds that are complementary in shape and charge to target proteins and as a consequence, bind the target proteins and function the therapeutic efficacy.

In addition, drugs are designed from traditional physiology-based treat exclusive target protein. The “Lock and Key” hypothesis was postulated to characterize the mechanism of enzyme activity. It states that the binding site has to have a unique geometric shape that is complementary to the substrate in order for them to bind. However, later experiments fail to validate this hypothesis due to the new discovery that enzyme binding is partially flexible rather than rigid. The “induced fit” hypothesis was introduced by Koshland [64] which assumes the docking between substrate and enzyme can be governed by Gibbs binding free energy. Plenty of docking algorithms emerged based on the new hypothesis, and the main difference among them is the scoring functions that rank the docking likelihood. DOCK treats the protein as a rigid body whereas ligands as flexible bodies, and utilizes three types of scoring functions to rank the ligands: shape feature, electrostatic potential and force field value. AUTODOCK utilizes the interaction energy which is calculated from atomic affinity potentials combined with electrostatic potential grid for the ranking. FlexX’s scoring function accounts for buried surface area, flexibility of the ligand, hydrogen bonds, non-polar interactions, salt bridges and also both enthalpic and entropic information.

These methods described above are called structure-based targeting prediction. It relies on the three dimensional structure of the target protein, and the structure is usually derived from x- ray crystallography or NMR spectroscopy. However, the speed of sequence discovery is far beyond over the speed of structure identification, and it results in the appearance of homology modeling methods to build target protein based on experimental knowledge. However, the error from homology modeling has to be taken into account during the target prediction. The other method is ligand-based targeting prediction, so called pharmacophore-based prediction. A pharmacophore is defined as a template that describes the essential features of a molecule. Ligand-based targeting focuses on the comparison between pharmacophore and ligand moieties. The scoring function screens top ligands in comparison, and these ligands have the propensity to bind the target protein.

An accurate and efficient method for protein surface characterization and comparison can play an important role in rational drug design. For example, analysis of protein surfaces could help identify protein binding pockets so that the requirements for a given pharmaceutical compound's size and binding orientation can be determined. Furthermore, knowledge of the protein conformation helps researchers develop specific pharmaceuticals for a given disease. This analysis can also assist in the investigation of protein-protein interactions and give researchers insight into the biological processes of the cell. We propose to develop a drug-target identification system that relies on the 2-D surface signature model. For a given drug with at least one known binding site, we will search for other protein structures containing similar surface patches, and rank the results by their similarity scores. An advantage of our proposed method over other existing off-

target prediction methods is its ability to not only predict potential targets, but also identify the binding sites.

Benchmark test

The benchmark dataset is from Metapocket [65]. The dataset contains 198 drug-target complexes with a 40% sequence similarity threshold. We select three ligand groups that contain at least 10 proteins in the dataset: Adenine (ADE, DB00173), Glutathione (GSH, DB00143), and Pyridoxal Phosphate (PLP, DB00114). The total number of proteins is 35: 10 proteins for group ADE, 10 proteins for group GSH, and 15 proteins for group PLP. Since the ligand information is provided in the dataset, the binding site can be easily calculated based on point distance between surface point and ligand. Thus, the binding site database is constructed, containing 35 binding sites, including binding site surface image, ligand group and color information.

A query protein, with unknown binding site knowledge, is searched against the database, in other words, all the binding sites in the database are compared against it, and a similarity score vector is maintained, along with corresponding predicted binding site locations for that query protein. Query protein is considered to bind the same ligand with the most similar binding site. For the benchmark testing, ligand group ADE is selected for evaluation of prediction accuracy. Proteins in ADE are assumed with unknown binding site information and are searched against the database excluding the binding site from the same protein. This gives a vector of similarity score with size 34. For example, protein (pdb: 1d2a) is compared with all the proteins in the database except the binding site generated from protein 1d2a itself. The result contains a score vector of size 34. Proteins in the same ligand group are expected to be more similar than binding sites from

other ligand groups. Since protein 1d2a binds with ligand ADE, the total 9 binding sites in the database that bind ADE are expected to locate at the first 9 positions in the score vector.

A true positive number is defined as the number of binding sites that are predicted to bind the same ligand actually bind the same ligand. A false positive number is defined as the number of binding sites that are predicted to bind the same ligand actually bind a different ligand.

Color optimization

As described in the previous section, color code and color weight are essential in binding site detection in that the detection is strongly dependent on the surface features. In our application, we use grid walk to tune the color parameters. Grid walk is a method of hyper parameter tuning algorithms. It is like the grid search, which evaluates the cost function at vertices of a grid, and selects the parameter settings from this grid that have the lowest function cost. But unlike grid search that transverses the whole grid, grid walk starts from an initial vertex, only evaluates the nearby vertices and moves the next vertex which has a lower cost. The nearby vertices are chosen according to the step in grid walk, as the algorithm moves closer toward the goal, the step decreases. For example, the starting point is at location (x,y) and if the step is defined as 5, the nearby vertices are ([x-5,y-5], [x-5,y+5], [x+5,y-5], [x+5,y+5]), in both positive and negative directions. The step decreases when it is getting closer to the best settings, and when the step size is too small to make further progress, the algorithm stops and the best parameters are found.

In the application, the cost function is defined as the square root difference between two images, the smaller the cost is, and the better the parameters are. Protein pairs that show

similarities are provided as the test datasets in the optimization, and the initial color channel is [1,1,1], and grid walk is able to calculate the cost function at the starting point and moves to the best color channel weight for each protein pairs.

Results

The ROC curve is plotted in Figure 36 for different color code, and AUC score is shown in Table 4. The color code numbers in the figure represent color codes of in Table 2 in order. The best AUC is 0.79 and the worst is 0.63. It is shown that color code EHC (3) outperforms the rest, indicating the color combination of electrostatic potential, hydrophobicity and evolution are major contributions to the prediction. In color code CHV (2), where electrostatic potential is removed, template matching performs the worst. This also indicates for small molecules bind with ligand ADE, the correct binding site detection depends largely on the electrostatic potential information.

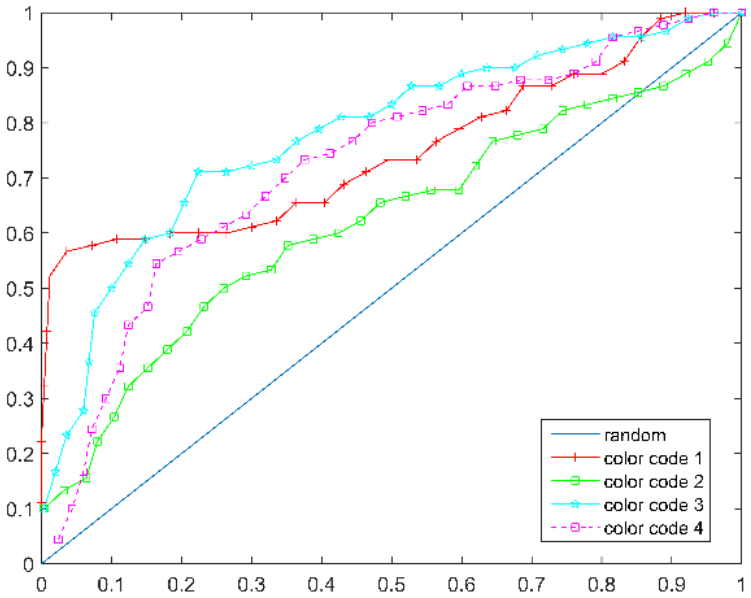

Figure 36 ROC curve for binding site prediction.

Table 4 AUC for different color codes

Color code	AUC
EHV	0.76
CHV	0.63
ECV	0.79
EHC	0.74

We have listed some protein examples from the prediction, shown in Figure 37, Figure 38 and Figure 40. The name of the proteins are shown under each image

respectively. Figure 37 shows prediction in color coder CHV, which is the combination of evolution conservation, hydrophobicity and curvature. The first column is the query (unknown) proteins. The query protein is segmented into multiple sub sections based on general segmentation, and each of the section is compared against the database. The first column shows the predicted section of the query protein. The red rectangle is the actual binding site location for the query protein, and the green rectangle is the predicted binding site location detected based on the second image. The second image is the binding site image that is voted top 3 in the database that have highest similarities with query protein. The red rectangles on second column are the template windows for each protein in the database, and are also the binding site region of the proteins in the database. Note that the size of the green rectangle (predicted) matches with the size of the red rectangle (template window) on the second column. This is because the actual size of the binding site of query protein is unknown, and the template window is used to predict the size of the binding site of query protein.

For the binding site images (second column), one can see the high conservation value (red channel) around the binding site region. Protein 1q8y also shows a strong hydrophobicity on the binding site region that leads a yellow patch on the area. The query protein 1cb0 also has the similar yellow patch around binding site region of itself. In contrast, binding site region of protein 1bj4 has a redder color pattern, and the query protein 1d2a shares more similar color on the binding site region compared to the rest of green area.

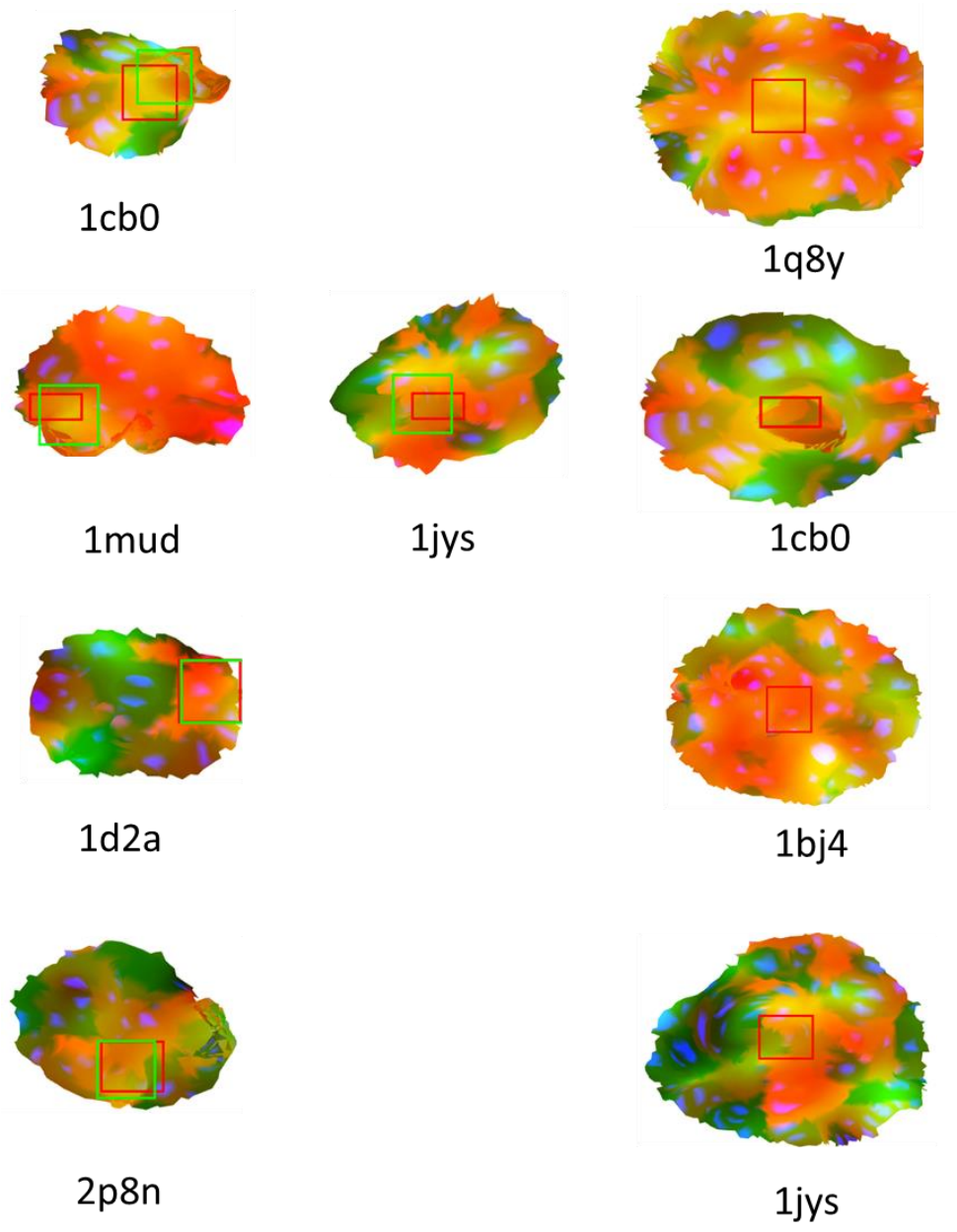

Figure 37 Template matching results using color code CHV

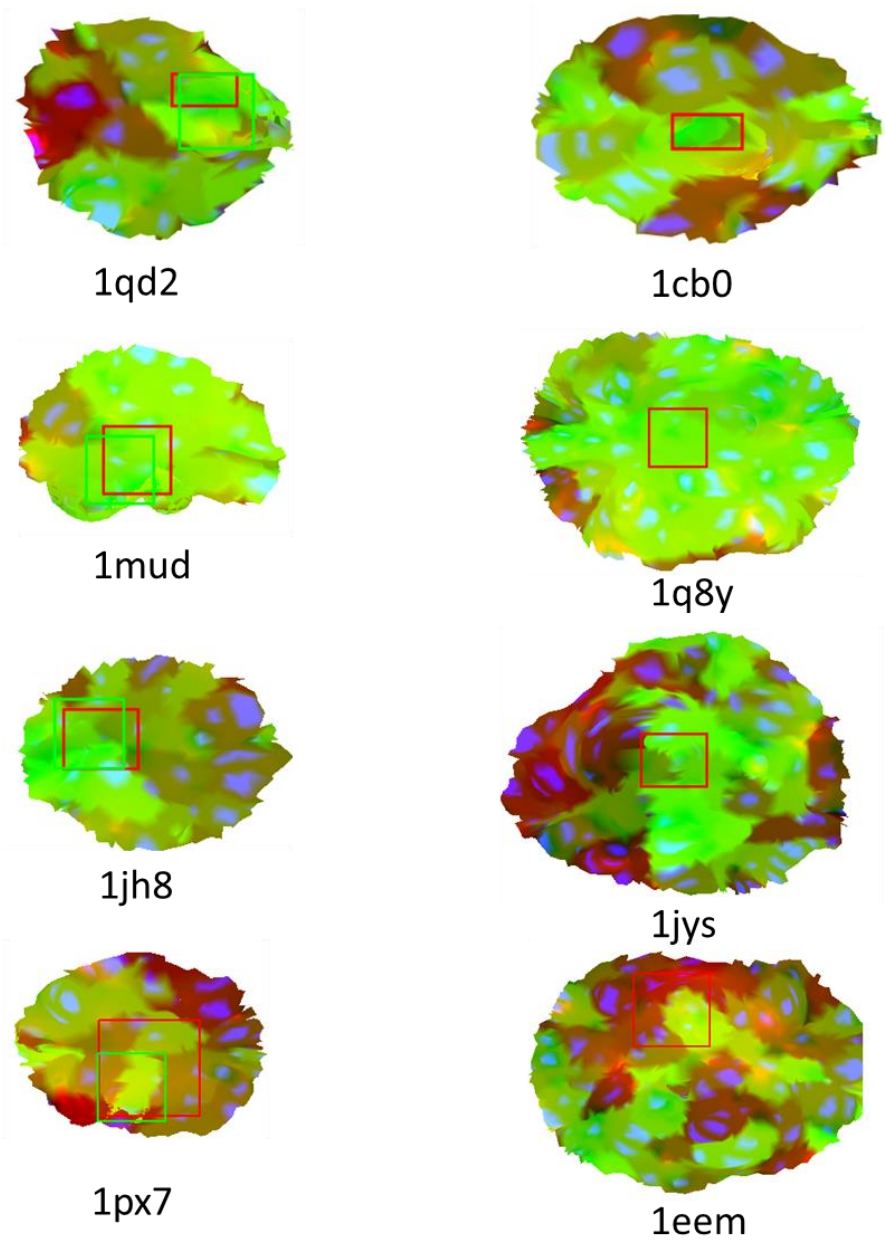

Figure 38 Template matching results using color code ECV

Figure 38 shows the template matching results using color code ECV, a combination of electrostatic potential, evolutionary conservation, and curvature. The first three pairs bind with ligand ADE, and the last pair bind with PLP. It is clear to see for proteins that bind

with ADE, when evolutionary conservation takes up the green channel, the binding site images show the strong green color, especially around binding site region. This means amino acids at binding sites are highly conserved and evolutionary conservation is the dominant feature, and the electrostatic potential and curvature have relatively low value. In contrast, the last protein pair, protein 1px7 and 1eem, which bind with PLP do not have the large conservation area, but only a small fraction on image around the binding site (bright yellow area) appears high conservation value but also strong electrostatic potential. Figure 39 shows the 3-D structures of the two proteins with color code ECV. The binding site regions are covered with bright yellow as compared to the neighboring area. This also demonstrates that for different ligand binding sites, there are variations in feature values, and the conservation areas locate differently.

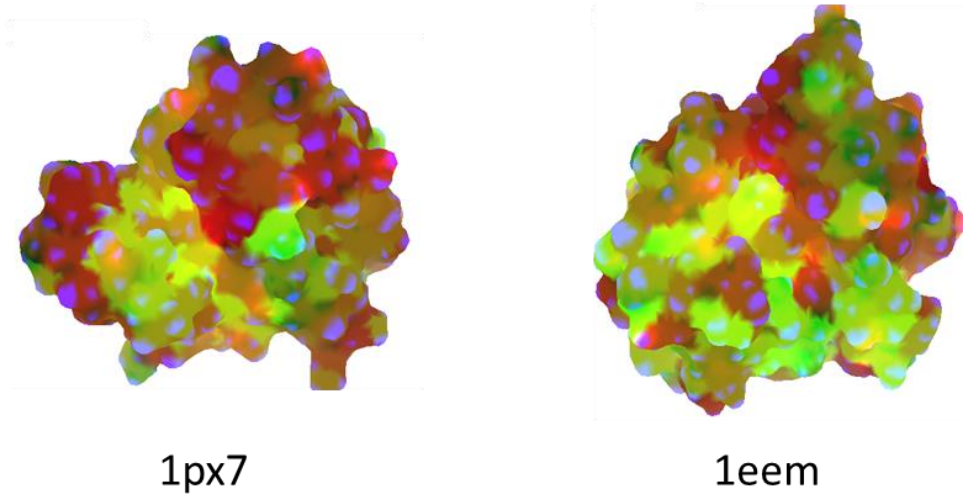

Figure 39 Color code ECV applied on surfaces of proteins 1px7 and 1eem.

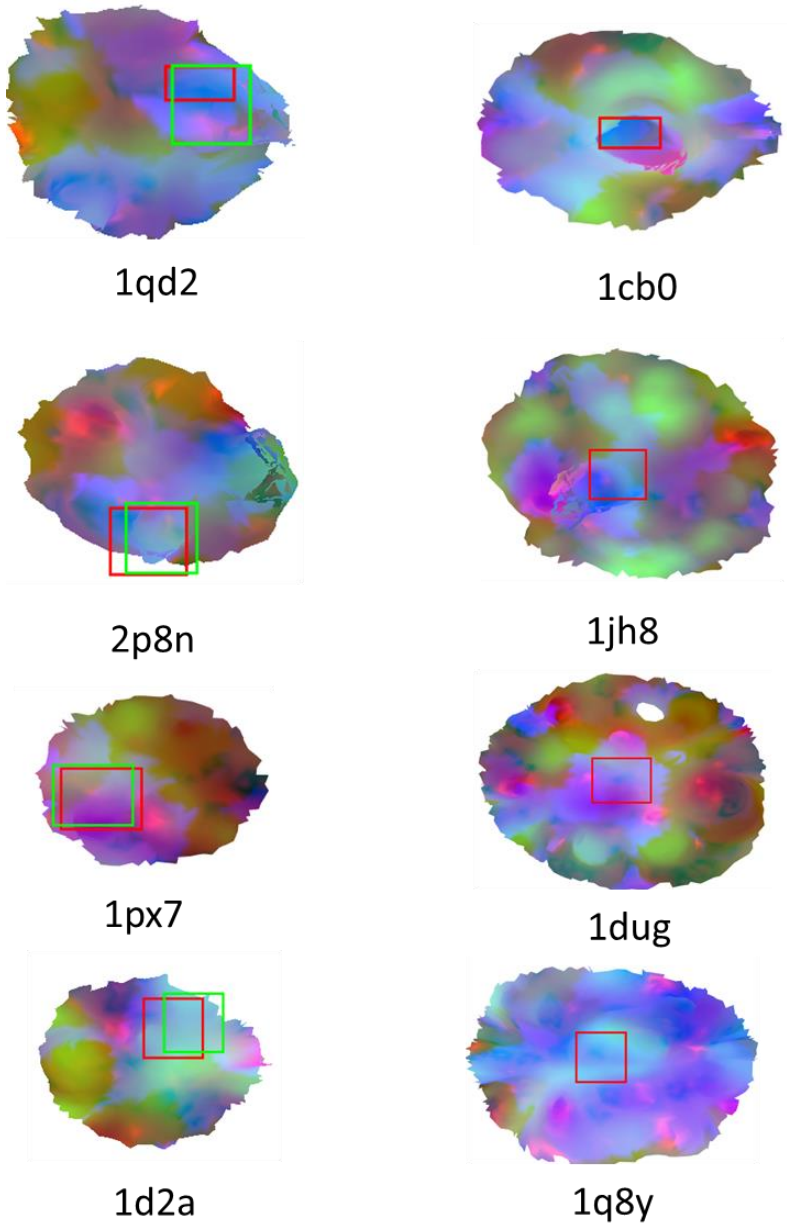

Figure 40 Template matching results using color code EHC

Figure 40 shows the prediction results using color code EHC, which has electrostatic potential, hydrophobicity and evolution conservation. As the above three color codes demonstrate, the prediction results perform differently, meaning that, proteins can be

detected by color code EHC may not be detected by color code ECV, and vice versa. This is due to the different color features in proteins. Thus, choosing the correct color combination is essential in our application for binding site detection. Not only the color code, but also the color weights in template matching. Initially, the color weights in the template matching for RGB channels is given equal amount, for example, the RGB channels are a vector of [1,1,1]. However, color channels can contribute unequally to the template matching, and it is possible that one channel contributes its information more than the rest of the colors. Thus, it is important to find the best color channel weight for each template matching using color optimization. The color optimization for color channel weights is discussed in previous section.

Time complexity

For the time consumption of one protein, we choose protein 1y59 as a general example. Protein 1y59 contains 8502 surface point, and 17010 triangles. The time consumption for processing an example protein 1y59 is in Table 5. By using the parallel computing, we can complete the computation for multiple sections at the same time. Template matching algorithm takes around 2 seconds for matching one template with one image. For a protein with 100 sub sections, template matching takes around 3 minutes. We cache the best similarity score for each protein pair. The time burden of this algorithm lies in Dimension Reduction methods. We selected 5 random proteins from above drug-target dataset. The average surface number is 9000, and the average time consumption for dimension reduction is 23 minutes. However, this can be solved via pre-calculation. Our algorithm does not calculate dimension

reduction during the running time. Instead, it calculates and caches it into database beforehand. Thus, when there are enough pre-calculations in the database, the only time complexity of the running time is dependent on the speed of template matching.

Table 5 Time consumption for protein 1y59

Pdb:1y59	Time
Surface calculation	2.098s
Electrostatic &hydrophobicity	23.2s
Conservation	15min
Segmentation	20.2s
Dimension reduction	21min

Chapter 8: Conclusion

Protein function annotation has always been a scientific challenging, and many researches have been done to improve the annotation accuracy and prediction rate. However, no matter using pure geometry information for pocket identification or taking advantage of physicochemical information, there is no current method that is satisfying for the accurate function annotation, and more underground biological knowledge are needed. We propose our method with a combination of informative geometry and physicochemical features, hope to make functional inferences of proteins, especially for distantly related proteins. Protein structure analysis has traditionally depended on backbone-only analysis, which is not amenable to detailed characterization of the protein, especially for the drug-design applications. Our dimension reduction approach makes the surface-based characterization of proteins computationally feasible and promises to be a powerful representation of protein structures.

Unlike protein structure alignment methods, which focus mainly on the global geometric similarity between two proteins, this method predicts the protein functional sites by a novel representation and comparison method for protein surface analysis.

Our method can capture various surface features in additional geometrical similarity, such as hydrophobicity, evolutionary conservation and electrostatic potential, in a computationally efficient way. Another advantage of our methods over protein structure alignment is that we focus on the surface directly and thus be able to detect the local similarities between two proteins.

There are several future works we would like to discuss and address. Because of the complex protein structures, especially proteins with long and narrow binding sites,

binding site segmentation is not able to capture the complete binding site since the cavity is buried deeper than the center of the mass where it is cut. The default solution is to a manually void cutting the binding site. Proteins that have long and narrow binding tunnels will have cutting parameters manually adjusted larger to avoid cutting on the binding site since the tunnel is buried deeper than the center of the mass. Image erosion and dilation are then used to cover the hollow caused by the cutting. However, the future work could involve automatic detection of this type of binding site, and adjust the cutting height and angle automatically. One of the limitations of our application is the time limitation. For the surface analysis and comparison, we recommend to use super computers or the more powerful computers to increase the speed. We intend to perform on a larger scale dataset, for example, computing on the entire PDB, and developing a surface based binding site database. However, this should also be computed on more powerful super computers, because the computation for each protein is time consuming, and there are over 100,000 structures in the PDB. For the weighted features mentioned above in Color optimization. It can be performed in more details, and weighted signatures can be developed for each protein family by using machine learning techniques.

Appendix

The dataset is from metapocket [65], which is derived from DrugPort, DrugBank. This dataset contains 198 drug-target complexes and the similarity threshold is 40%.

pdbid(chainID) protein description (drug)

1azm(A) Carbonic anhydrase Acetazolamide(AZM,DB00819)

1jd0(A) Carbonic anhydrase Acetazolamide(AZM,DB00819)

1fwe(C) Urease, beta-subunit Acetohydroxamic Acid(HAE,DB00551)

3ba0(A) Macrophage metalloelastase Acetohydroxamic Acid(HAE,DB00551)

1cb0(A) 5'-deoxy-5'-methylthioadenosine phosphorylase Adenine(ADE,DB00173)

1d2a(B) Tyrosine phosphatase Adenine(ADE,DB00173)

1hqc(A) Holliday junction helicase RuvB Adenine(ADE,DB00173)

1jh8(A) Nicotinate mononucleotide:5,6-dimethylbenzimidazole phosphoribosyltransferase (CobT) Adenine(ADE,DB00173)

1jys(A) 5'-Methylthioadenosine/S-Adenosylhomocysteine nucleosidase Adenine(ADE,DB00173)

1lpd(A) Dianthin 30 Adenine(ADE,DB00173)

1lu1(A) Legume lectin Adenine(ADE,DB00173)

1mud(A) Catalytic domain of MutY Adenine(ADE,DB00173)

1od2(B) Acetyl-coenzyme A carboxylase Adenine(ADE,DB00173)

1q8y(B) Sky1p Adenine(ADE,DB00173)

1qb7(A) Adenine PRTase Adenine(ADE,DB00173)

1qci(A) Pokeweed antiviral protein alpha Adenine(ADE,DB00173)

1qd2(A) alpha-Trichosanthin Adenine(ADE,DB00173)

1s2d(A) Purine transdeoxyribosylase Adenine(ADE,DB00173)

1xe8(A) Hypothetical protein YML079W Adenine(ADE,DB00173)

1yxm(B) Peroxisomal trans 2-enoyl CoA reductase Adenine(ADE,DB00173)

1zn7(A) Adenine PRTase Adenine(ADE,DB00173)

2p8n(A) Ricin A-chain Adenine(ADE,DB00173)

2nvu(B) UBA3 Adenosine triphosphate(ATP,DB00171)

3iyt(A) Apoptotic protease-activating factor 1 Adenosine triphosphate(ATP,DB00171)

1yhm(A) Farnesyl pyrophosphate synthase Alendronate(AHD,DB00630)

2f92(F) Farnesyl diphosphate synthase Alendronate(AHD,DB00630)

1f5l(A) Urokinase-type plasminogen activator (LMW U-PA), catalytic domain Amiloride(AMR,DB00594)

1cea(A) Plasminogen Aminocaproic Acid(ACA,DB00513)

1pk2(A) Plasminogen Aminocaproic Acid(ACA,DB00513)

1pbc(A) p-Hydroxybenzoate hydroxylase, PHBH Aminosalicylic Acid(BHA,DB00233)

1sxk(A) Snake phospholipase A2 Aminosalicylic Acid(BHA,DB00233)

2aou(A) Histamine methyltransferase Amodiaquine(CQA,DB00613)

1oxr(A) Snake phospholipase A2 Aspirin(AIN,DB00945)

1tgm(A) Snake phospholipase A2 Aspirin(AIN,DB00945)

1hwk(A) NAD-binding domain of HMG-CoA reductase Atorvastatin(117,DB01076)

1th6(A) Snake phospholipase A2 Atropine(OIN,DB00572)

1tuf(A) Diaminopimelate decarboxylase LysA Azelaic Acid(AZ1,DB00548)

1lxf(C) Troponin C Bepridil(BEP,DB01244)

2bdm(A) Mammalian cytochrome p450 2b4 Bifonazole(TMI,DB04794)

1s19(A) Vitamin D nuclear receptor Calcipotriol(MC9,DB02300)

1b3n(A) Beta-ketoacyl-ACP synthase II Cerulenin(CER,DB01034)

1fj8(A) Beta-ketoacyl-ACP synthase I Cerulenin(CER,DB01034)

1qhy(A) Chloramphenicol phosphotransferase Chloramphenicol(CLM,DB00446)

1usq(A) DraA/Afimbrial adhesin Afa-III Chloramphenicol(CLM,DB00446)

2xat(A) Xenobiotic acetyltransferase Chloramphenicol(CLM,DB00446)

3cla(A) Chloramphenicol acetyltransferase, CAT Chloramphenicol(CLM,DB00446)

1p0m(A) Butyryl cholinesterase Choline(CHT,DB00122)

2fy3(A) Choline o-acetyltransferase Choline(CHT,DB00122)

1itu(A) Renal dipeptidase Cilastatin(CIL,DB01597)

1fcm(A) AMPC beta-Lactamase, class C Cloxacillin(CXN,DB01147)

3b6r(B) Creatine kinase b-type Creatine(CRN,DB00148)

1n2z(A) Vitamin B12 binding protein BtuF Cyanocobalamin(CNC,DB00115)

2gsk(A) TonB Cyanocobalamin(CNC,DB00115)

3h6t(A) Glutamate receptor 2 Cyclothiazide(CYZ,DB00606)

1mrl(A)	Xenobiotic acetyltransferase	Dalfopristin(DOL,DB01764)
1m2z(A)	Glucocorticoid receptor	Dexamethasone(DEX,DB01234)
3cfq(A)	Transthyretin	Diclofenac(DIF,DB00586)
1s9p(A)	Orphan nuclear receptor ERR3	Diethylstilbestrol(DES,DB00255)
1tt6(A)	Transthyretin (synonym: prealbumin)	Diethylstilbestrol(DES,DB00255)
3erd(A)	Estrogen receptor alpha	Diethylstilbestrol(DES,DB00255)
1c1p(A)	Trypsin(ogen)	Dimethyl sulfoxide(DMS,DB01093)
1dp0(A)	beta-Galactosidase, domains 2 and 4	Dimethyl sulfoxide(DMS,DB01093)
1kmv(A)	Dihydrofolate reductases, eukaryotic type	Dimethyl sulfoxide(DMS,DB01093)
1lj5(A)	Plasminogen activator inhibitor-1	Dimethyl sulfoxide(DMS,DB01093)
1ltq(A)	Polynucleotide kinase, phosphatase domain	Dimethyl sulfoxide(DMS,DB01093)
1q1c(A)	FKBP52, N-terminal domains	Dimethyl sulfoxide(DMS,DB01093)
1r6n(A)	E2 regulatory, transactivation domain	Dimethyl sulfoxide(DMS,DB01093)
1s2a(A)	Prostaglandin d2 11-ketoreductase (akr1c3)	Dimethyl sulfoxide(DMS,DB01093)
1tpf(A)	Triosephosphate isomerase	Dimethyl sulfoxide(DMS,DB01093)
1tz8(A)	Transthyretin (synonym: prealbumin)	Dimethyl sulfoxide(DMS,DB01093)
1yki(B)	Oxygen-insensitive NAD(P)H nitroreductase	Dimethyl sulfoxide(DMS,DB01093)
2biu(X)	Mitochondrial peptidyl-prolyl cis-trans isomerase, cyclophilin F	Dimethyl sulfoxide(DMS,DB01093)
2jhf(A)	Alcohol dehydrogenase	Dimethyl sulfoxide(DMS,DB01093)
2of1(A)	Staphylococcal thermonuclease	Dimethyl sulfoxide(DMS,DB01093)
3k4v(A)	Hiv-1 protease	Dimethyl sulfoxide(DMS,DB01093)
PDBID(chainID)	Description of protein	Drug(s)
3pah(A)	Phenylalanine hydroxylase, PAH	Epinephrine(ALE,DB00668)
1m17(A)	EGF receptor tyrosine kinase, Erbb-1	Erlotinib(AQ4,DB00530)
1x8v(A)	Cytochrome p450 14 alpha-sterol demethylase (cyp51)	Estriol(ESL,DB04573)
3dgq(A)	Glutathione s-transferase p	Ethacrynic acid(EAA,DB00903)
1h7x(A)	Dihydropyrimidine dehydrogenase, N-terminal domain	Fluorouracil(URF,DB00544)
1upf(D)	Uracil PRTase, Upp	Fluorouracil(URF,DB00544)
3kvv(A)	Uridine phosphorylase	Fluorouracil(URF,DB00544)
1uae(A)	UDP-N-acetylglucosamine enolpyruvyl transferase (EPT, MurA, MurZ)	Fosfomycin(FCN,DB00828)

1qca(A) Chloramphenicol acetyltransferase, CAT Fusidic Acid(FUA,DB02703)

2coj(A) Branched chain aminotransferase 1, cytosolic Gabapentin(GBN,DB00996)

1w6r(A) Acetylcholinesterase Galantamine(GNT,DB00674)

1ki2(A) Thymidine kinase Ganciclovir(GA2,DB01004)

1dug(A) Class alpha GST Glutathione(GSH,DB00143)

1eem(A) Class omega GST Glutathione(GSH,DB00143)

1f3a(A) Class alpha GST Glutathione(GSH,DB00143)

1fw1(A) Class zeta GST Glutathione(GSH,DB00143)

1px7(A) Class pi GST Glutathione(GSH,DB00143)

1q8m(A) TREM-1 (triggering receptor expressed on myeloid cells 1) Glutathione(GSH,DB00143)

1qgj(A) Plant peroxidase Glutathione(GSH,DB00143)

1r4w(A) Mitochondrial class kappa glutathione S-transferase Glutathione(GSH,DB00143)

2hgs(A) Eukaryotic glutathione synthetase, substrate-binding domain Glutathione(GSH,DB00143)

2ht9(A) Glutaredoxin-2 Glutathione(GSH,DB00143)

2vcq(A) Class sigma GST Glutathione(GSH,DB00143)

3dk8(A) Glutathione reductase Glutathione(GSH,DB00143)

2zt7(A) Glycyl-trna synthetase Glycine(GLY,DB00145)

5jdw(A) L-arginine: glycine amidinotransferase Glycine(GLY,DB00145)

1jxm(A) Psd-95 Guanidine(GAI,DB00536)

1kp3(A) Argininosuccinate synthetase, N-terminal domain Guanidine(GAI,DB00536)

1rbw(A) Ribonuclease A (also ribonuclease B, S) Guanidine(GAI,DB00536)

1umj(A) Cut A1 Guanidine(GAI,DB00536)

1xcl(A) Guanidinoacetate methyltransferase Guanidine(GAI,DB00536)

2cev(A) Arginase Guanidine(GAI,DB00536)

2f2q(A) Phage T4 lysozyme Guanidine(GAI,DB00536)

3inj(A) Aldehyde dehydrogenase, mitochondrial Guanidine(GAI,DB00536)

2bxg(A) Serum albumin Ibuprofen(IBP,DB01050)

3gwx(A) Peroxisome proliferator-activated receptor delta, PPAR-DELTA Icosapent(EPA,DB00159)

2bxn(A) Serum albumin Iodipamide(IDB,DB04711)

1w6f(A) Arylamine N-acetyltransferase Isoniazid(ISZ,DB00951)

1nsi(A)	Nitric oxide (NO) synthase oxygenase domain	L-Arginine(ARG,DB00125)
2nz2(A)	Argininosuccinate synthase	L-Aspartic Acid(ASP,DB00128),L-Citrulline(CIR,DB00155)
2jai(A)	Ng, ng-dimethylarginine dimethylaminohydrolase 1	L-Citrulline(CIR,DB00155)
6pah(A)	Phenylalanine hydroxylase, PAH	Levodopa(DAH,DB01235)
2c6g(A)	Glutamate carboxypeptidase II	L-Glutamic Acid(GLU,DB00142)
2xhd(A)	Glutamate receptor 2	L-Glutamic Acid(GLU,DB00142)
3czd(A)	Glutaminase kidney isoform	L-Glutamic Acid(GLU,DB00142)
2h79(A)	Thra protein	Liothyronine(T3,DB00279)
2c6n(A)	Angiotensin-converting enzyme, somatic isoform	Lisinopril(LPR,DB00722)
3bju(A)	Lysyl-trna synthetase	L-Lysine(LYS,DB00123)
1rv7(B)	Human immunodeficiency virus type 1 protease	Lopinavir(AB1,DB01601)
3gmz(A)	Arginase-1	L-Ornithine(ORN,DB00129)
3jdw(A)	L-arginine: glycine amidinotransferase	L-Ornithine(ORN,DB00129)
1wap(A)	Trp RNA-binding attenuation protein (TRAP)	L-Tryptophan(TRP,DB00150)
2azx(A)	Tryptophanyl-tRNA synthetase (TrpRS)	L-Tryptophan(TRP,DB00150)
2qr2(B)	Quinone reductase type 2 (menadione reductase)	Menadione(VK3,DB00170)
1z11(A)	Cytochrome p450, family 2, subfamily a,polypeptide 6	Methoxsalen(8MO,DB00553)
1w0g(A)	Mammalian cytochrome P450 3a4	Metyrapone(MYT,DB01011)
2w8y(A)	Progesterone receptor	Mifepristone(486,DB00834)
1ffy(A)	Isoleucyl-tRNA synthetase (IleRS)	Mupirocin(MRC,DB00410)
1jr1(A)	Inosine monophosphate dehydrogenase (IMPDH)	Mycophenolic acid(MOA,DB01024)
1me7(A)	Inosine monophosphate dehydrogenase (IMPDH)	Mycophenolic acid(MOA,DB01024)
2agd(A)	Beta-1,4-galactosyltransferase 1	N-Acetyl-D-glucosamine(NAG,DB00141)
2ch5(A)	N-acetylglucosamine kinase, NAGK	N-Acetyl-D-glucosamine(NAG,DB00141)
1f17(A)	Short chain L-3-hydroxyacyl CoA dehydrogenase	NADH(NAI,DB00157)
1i0z(A)	Lactate dehydrogenase	NADH(NAI,DB00157)
1pj2(A)	Mitochondrial NAD(P)-dependent malic enzyme	NADH(NAI,DB00157)
1zmd(A)	Dihydrolipoyl dehydrogenase	NADH(NAI,DB00157)
2j6l(A)	Aldehyde dehydrogenase family 7 member a1	NADH(NAI,DB00157)
1p7r(A)	Cytochrome P450-CAM	Nicotine(NCT,DB00184)

1uw6(A) Acetylcholine binding protein (ACHBP) Nicotine(NCT,DB00184)

1td7(A) Snake phospholipase A2 Niflumic Acid(NFL,DB04552)

4pah(A) Phenylalanine hydroxylase, PAH Norepinephrine(LNR,DB00368)

1sqn(A) Progesterone receptor Norethindrone(NDR,DB00717)

3d90(A) Progesterone receptor Norgestrel(NOG,DB00506)

2f89(F) Farnesyl diphosphate synthase Pamidronate(210,DB00282)

1ju6(A) Thymidylate synthase Pemetrexed(LYA,DB00642)

1gtb(A) Class alpha GST Praziquantel(PZQ,DB01058)

1e7a(A) Serum albumin Propofol(PFL,DB00818)

1bj4(A) Serine hydroxymethyltransferase Pyridoxal Phosphate(PLP,DB00114)

1fa9(A) Glycogen phosphorylase Pyridoxal Phosphate(PLP,DB00114)

1h0c(A) Alanine-glyoxylate aminotransferase Pyridoxal Phosphate(PLP,DB00114)

1m54(A) Cystathionine beta-synthase Pyridoxal Phosphate(PLP,DB00114)

1nrg(A) Pyridoxine 5'-phoshate oxidase (PNP oxidase) Pyridoxal Phosphate(PLP,DB00114)

1oat(A) Ornithine aminotransferase Pyridoxal Phosphate(PLP,DB00114)

1p5j(A) L-serine dehydratase Pyridoxal Phosphate(PLP,DB00114)

2abj(A) Branched-chain-amino-acid aminotransferase,cytosolic Pyridoxal Phosphate(PLP,DB00114)

2cft(A) Pyridoxal phosphate phosphatase Pyridoxal Phosphate(PLP,DB00114)

2hzp(A) Kynureninase Pyridoxal Phosphate(PLP,DB00114)

2jis(A) Cysteine sulfinic acid decarboxylase Pyridoxal Phosphate(PLP,DB00114)

2oo0(A) Ornithine decarboxylase Pyridoxal Phosphate(PLP,DB00114)

2xh1(A) Kynurenine/alpha-aminoadipate aminotransferase,mitochondrial Pyridoxal Phosphate(PLP,DB00114)

3cog(A) Cystathionine gamma-lyase Pyridoxal Phosphate(PLP,DB00114)

3dyd(A) Tyrosine aminotransferase Pyridoxal Phosphate(PLP,DB00114)

3e77(A) Phosphoserine aminotransferase Pyridoxal Phosphate(PLP,DB00114)

3ii0(A) Aspartate aminotransferase, cytoplasmic Pyridoxal Phosphate(PLP,DB00114)

3l6b(A) Serine racemase Pyridoxal Phosphate(PLP,DB00114)

3fhx(A) Pyridoxal kinase Pyridoxal(PXL,DB00147)

1j3j(A) Bifunctional enzyme dihydrofolate reductase-thymidylate synthase, DFR domain Pyrimethamine(CP6,DB00205)

3bg3(A) Pyruvate carboxylase, mitochondrial Pyruvic acid(PYR,DB00119)

2qxs(A) Estrogen receptor Raloxifene(RAL,DB00481)

3dzy(D) Peroxisome proliferator-activated receptor gamma Rosiglitazone(BRL,DB00412)

1hwl(A) NAD-binding domain of HMG-CoA reductase Rosuvastatin(FBI,DB01098)

2obv(A) S-adenosylmethionine synthetase isoform type-1 S-Adenosylmethionine(SAM,DB00118)

3b9m(A) Serum albumin Salicyclic acid(SAL,DB00936)

3c6m(A) Spermine synthase Spermine(SPM,DB00127)

2ab2(A) Mineralocorticoid receptor Spironolactone(SNL,DB00421)

1y4l(B) Snake phospholipase A2 Suramin(SVR,DB04786)

1y8e(A) Complement control protein Suramin(SVR,DB04786)

2h9t(H) ThrombinSuramin(SVR,DB04786)

2nyr(B) NAD-dependent deacetylase sirtuin-5 Suramin(SVR,DB04786)

1q6i(A) Peptidyl-prolyl cis-trans isomerase FkpA Tacrolimus(FK5,DB00864)

1udu(A) cGMP-specific 3',5'-cyclic phosphodiesterase pde5a1-Ibmx Tadalafil(CIA,DB00820)

3ert(A) Estrogen receptor alpha Tamoxifen(OHT,DB00675)

1kdk(A) Sex hormone-binding globulin Testosterone(DHT,DB00624)

1xj7(A) Androgen receptor Testosterone(DHT,DB00624)

1ig0(A) Thiamin pyrophosphokinase, substrate-binding domain Thiamine(VIB,DB00152)

1ig3(A) Thiamin pyrophosphokinase, substrate-binding domain Thiamine(VIB,DB00152)

1sbr(A) Putative thiamin/HMP-binding protein YkoF Thiamine(VIB,DB00152)

2vdm(B) Immunoglobulin heavy chain gamma constant domain 1, CH1-gamma Tirofiban(AGG,DB00775)

1ihi(A) 3-alpha-hydroxysteroid dehydrogenase Ursodeoxycholic acid(IU5,DB01586)

1pn3(A) TDP-epi-vancosaminyltransferase GtfA Vancomycin(DVV,DB00512)

1c3s(A) HDAC homologue Vorinostat(SHH,DB02546)

1t69(A) Histone deacetylase 8, HDAC8 Vorinostat(SHH,DB02546)

2f8z(F) Farnesyl diphosphate synthase Zoledronate(ZOL,DB00399)

Reference

1. Bernstein, F.C., et al., *The Protein Data Bank: a computer-based archival file for macromolecular structures.* J Mol Biol, 1977. **112**(3): p. 535-42.
2. Chruszcz, M., et al., *Unmet challenges of structural genomics.* Curr Opin Struct Biol, 2010. **20**(5): p. 587-97.
3. Holm, L. and C. Sander, *Protein structure comparison by alignment of distance matrices.* J Mol Biol, 1993. **233**(1): p. 123-38.
4. Grindley, H.M., et al., *Identification of tertiary structure resemblance in proteins using a maximal common subgraph isomorphism algorithm.* J Mol Biol, 1993. **229**(3): p. 707-21.
5. Nussinov, R. and H.J. Wolfson, *Efficient detection of three-dimensional structural motifs in biological macromolecules by computer vision techniques.* Proc Natl Acad Sci U S A, 1991. **88**(23): p. 10495-9.
6. May, A.C. and M.S. Johnson, *Improved genetic algorithm-based protein structure comparisons: pairwise and multiple superpositions.* Protein Eng, 1995. **8**(9): p. 873-82.
7. Sacan, A., I.H. Toroslu, and H. Ferhatosmanoglu, *Integrated search and alignment of protein structures.* Bioinformatics, 2008. **24**(24): p. 2872-9.
8. Nagano, N., C.A. Orengo, and J.M. Thornton, *One fold with many functions: the evolutionary relationships between TIM barrel families based on their sequences, structures and functions.* J Mol Biol, 2002. **321**(5): p. 741-65.
9. Polgar, L., *The catalytic triad of serine peptidases.* Cell Mol Life Sci, 2005. **62**(19-20): p. 2161-72.
10. Nadzirin, N., et al., *SPRITE and ASSAM: web servers for side chain 3D-motif searching in protein structures.* Nucleic Acids Res, 2012. **40**(Web Server issue): p. W380-6.
11. Golovin, A., K. Henrick, and G. Kleywegt, *Integration of chemical information with protein sequences and 3D structures.* Journal of Cheminformatics, 2010. **2**(Suppl 1): p. 1-1.
12. Sacan, A., et al., *LFM-Pro: a tool for detecting significant local structural sites in proteins.* Bioinformatics, 2007. **23**(6): p. 709-16.
13. Hutchinson, E.G. and J.M. Thornton, *PROMOTIF--a program to identify and analyze structural motifs in proteins.* Protein Sci, 1996. **5**(2): p. 212-20.
14. Shulman-Peleg, A., R. Nussinov, and H.J. Wolfson, *Recognition of functional sites in protein structures.* J Mol Biol, 2004. **339**(3): p. 607-33.
15. Liang, J., H. Edelsbrunner, and C. Woodward, *Anatomy of protein pockets and cavities: measurement of binding site geometry and implications for ligand design.* Protein Sci, 1998. **7**(9): p. 1884-97.
16. Laskowski, R.A., et al., *Protein clefts in molecular recognition and function.* Protein Sci, 1996. **5**(12): p. 2438-52.
17. Laurie, A.T. and R.M. Jackson, *Q-SiteFinder: an energy-based method for the prediction of protein-ligand binding sites.* Bioinformatics, 2005. **21**(9): p. 1908-16.

18. Hendlich, M., F. Rippmann, and G. Barnickel, *LIGSITE: automatic and efficient detection of potential small molecule-binding sites in proteins.* J Mol Graph Model, 1997. **15**(6): p. 359-63, 389.
19. Sael, L., et al., *Rapid comparison of properties on protein surface.* Proteins, 2008. **73**(1): p. 1-10.
20. Das, S., A. Kokardekar, and C.M. Breneman, *Rapid comparison of protein binding site surfaces with property encoded shape distributions.* J Chem Inf Model, 2009. **49**(12): p. 2863-72.
21. Najmanovich, R., N. Kurbatova, and J. Thornton, *Detection of 3D atomic similarities and their use in the discrimination of small molecule protein-binding sites.* Bioinformatics, 2008. **24**(16): p. i105-11.
22. Fanning, D.W., J.A. Smith, and G.D. Rose, *Molecular Cartography of Globular-Proteins with Application to Antigenic Sites.* Biopolymers, 1986. **25**(5): p. 863-883.
23. Pawlowski, K. and A. Godzik, *Surface Map Comparison: Studying Function Diversity of Homologous Proteins.* Journal of Molecular Biology, 2001. **309**(3): p. 793 - 806.
24. Lee, B. and F.M. Richards, *The interpretation of protein structures: estimation of static accessibility.* J Mol Biol, 1971. **55**(3): p. 379-400.
25. Richards, F.M., *Areas, volumes, packing and protein structure.* Annu Rev Biophys Bioeng, 1977. **6**: p. 151-176.
26. J, G. and L.B. B, *Macromolecular shape and surface maps by solvent exclusion.* Proc Natl Acad Sci U S A, 1978 January;. **75**: p. 303-307.
27. Delano, W.L., *The PyMOL Molecular Graphics System,Version 1.2r3pre, Schrödinger, LLC.* 2002.
28. Jmol. *Jmol:an open-source Java viewer for chemical structures in 3D*. Available from: http://www.jmol.org/.
29. Kitchen, D.B., et al., *Docking and scoring in virtual screening for drug discovery: methods and applications.* Nature Reviews Drug Discovery, 2004: p. 935-949.
30. Binkowski, T.A., A. Joachimiak, and J. Liang, *Protein surface analysis for function annotation in high-throughput structural genomics pipeline.* Protein Sci, 2005. **14**(12): p. 2972-81.
31. Connolly, M.L., *Analytical Molecular-Surface Calculation.* Journal of Applied Crystallography, 1983. **16**(Oct): p. 548-558.
32. Connolly, M.L., *Molecular-Surface Triangulation.* Journal of Applied Crystallography, 1985. **18**(Dec): p. 499-505.
33. Sanner, M.F., A.J. Olson, and J.C. Spehner, *Fast and robust computation of molecular surfaces.* Proceedings of the eleventh annual symposium on Computational geometry, 1995: p. 406-407.
34. Staib, L.H. and J.S. Duncan, *Model-based deformable surface finding for medical images.* IEEE Trans Med Imaging, 1996. **15**(5): p. 720-31.
35. Bairoch, A., *The PROSITE dictionary of sites and patterns in proteins, its current status.* Nucleic Acids Res, 1993. **21**(13): p. 3097-103.
36. !!! INVALID CITATION !!!

37. Venkatraman, V., L. Sael, and D. Kihara, *Potential for Protein Surface Shape Analysis Using Spherical Harmonics and 3D Zernike Descriptors.* Cell Biochemistry and Biophysics, 2009. **54**: p. 23-32.
38. Fischer, D., et al., *Surface motifs by a computer vision technique: searches, detection, and implications for protein-ligand recognition.* Proteins, 1993. **16**(3): p. 278-92.
39. Neves-Petersen, M.T. and S.B. Petersen, *Protein electrostatics: a review of the equations and methods used to model electrostatic equations in biomolecules--applications in biotechnology.* Biotechnol Annu Rev, 2003. **9**: p. 315-95.
40. Kyte, J. and R.F. Doolittle, *A simple method for displaying the hydropathic character of a protein.* J Mol Biol, 1982. **157**(1): p. 105-32.
41. Edgar, R.C., *MUSCLE: a multiple sequence alignment method with reduced time and space complexity.* BMC Bioinformatics, 2004. **5**: p. 113.
42. Shatsky, M., R. Nussinov, and H.J. Wolfson, *Optimization of multiple-sequence alignment based on multiple-structure alignment.* Proteins, 2006. **62**(1): p. 209-17.
43. Via, A., et al., *Protein surface similarities: a survey of methods to describe and compare protein surfaces.* Cell Mol Life Sci., 2000. **57**: p. 1970-7.
44. Bork, P., C. Sander, and A. Valencia, *Convergent evolution of similar enzymatic function on different protein folds: the hexokinase, ribokinase, and galactokinase families of sugar kinases.* Protein Sci, 1993. **2**(1): p. 31-40.
45. Kauvar, L.M. and H.O. Villar, *Deciphering cryptic similarities in protein binding sites.* Current Opinion in Biotechnology, 1998. **9**(4): p. 390 - 394.
46. Russell, R.B., P.D. Sasieni, and M.J.E. Sternberg, *Supersites within superfolds. Binding site similarity in the absence of homology.* Journal of Molecular Biology, 1998. **282**(4): p. 903 - 918.
47. van der Maaten, L.J.P., *An Introduction to Dimensionality Reduction Using Matlab.* Technical Report 07-06, MICC-IKAT, Maastricht University, Maastricht, The Netherlands, 2007.
48. Tenenbaum, J.B., V. de Silva, and J.C. Langford, *A global geometric framework for nonlinear dimensionality reduction.* Science, 2000. **290**(5500): p. 2319-23.
49. Peyre, G., *Graph theory toolbox* 2007.
50. Zhao, W., et al., *Face recognition: A literature survey.* Acm Computing Surveys (CSUR), 2003. **35**(4): p. 399-458.
51. Bradski, G. and A. Kaehler, *Learning OpenCV: Computer vision with the OpenCV library*. 2008: O'reilly.
52. Bertolazzi, P., C. Guerra, and G. Liuzzi, *A global optimization algorithm for protein surface alignment.* BMC Bioinformatics, 2010. **11**: p. 488.
53. Angaran, S., et al., *MolLoc: a web tool for the local structural alignment of molecular surfaces.* Nucleic Acids Res, 2009. **37**(Web Server issue): p. W565-70.
54. Nicholls, A., K.A. Sharp, and B. Honig, *Protein folding and association: insights from the interfacial and thermodynamic properties of hydrocarbons.* Proteins, 1991. **11**(4): p. 281-96.
55. Steinkellner, G., et al., *VASCo: computation and visualization of annotated protein surface contacts.* BMC Bioinformatics, 2009. **10**: p. 32.

56. Sheinerman, F.B., R. Norel, and B. Honig, *Electrostatic aspects of protein-protein interactions.* Curr Opin Struct Biol, 2000. **10**(2): p. 153-9.
57. Li, L., et al., *DelPhi: a comprehensive suite for DelPhi software and associated resources.* BMC Biophys, 2012. **5**: p. 9.
58. Cao, Y. and L. Li, *Improved protein-ligand binding affinity prediction by using a curvature-dependent surface-area model.* Bioinformatics, 2014.
59. Chen, X. and F. Schmitt, *Intrinsic surface properties from surface triangulation*, in *Computer Vision — ECCV'92*, G. Sandini, Editor. 1992, Springer Berlin Heidelberg. p. 739-743.
60. Dong chen-shi, W.G.-Z., *Curvatures estimation on triangular mesh.* Journal of Zhejiang University SCIENCE, 2005.
61. Altschul, S.F., et al., *Gapped BLAST and PSI-BLAST: a new generation of protein database search programs.* Nucleic Acids Res, 1997. **25**(17): p. 3389-402.
62. Porter, C.T., G.J. Bartlett, and J.M. Thornton, *The Catalytic Site Atlas: a resource of catalytic sites and residues identified in enzymes using structural data.* Nucleic Acids Res, 2004. **32**(Database issue): p. D129-33.
63. Berman, H.M., et al., *The Protein Data Bank.* Acta Crystallogr D Biol Crystallogr, 2002. **58**(Pt 6 No 1): p. 899-907.
64. Koshland, D.E., *Application of a Theory of Enzyme Specificity to Protein Synthesis.* Proc Natl Acad Sci U S A, 1958. **44**(2): p. 98-104.
65. Huang, B., *MetaPocket: a meta approach to improve protein ligand binding site prediction.* OMICS, 2009. **13**(4): p. 325-30.